

Percept formation from neural populations in sensory decision-making tasks

Adrien Wohrer^{1,*}, Christian Machens²

1 Group for Neural Theory, INSERM U960, École Normale Supérieure, Paris, France

2 Champalimaud Neuroscience Program, Lisbon, Portugal

* E-mail: adrien.wohrer@ens.fr

Abstract

We study a standard linear readout model of perceptual integration from a population of sensory neurons. We show that the readout can be associated to a set of characteristic equations which summarize the joint trial-to-trial covariance structure of neural activities and animal percept. These characteristic equations implicitly determine the readout parameters that were used by the animal to create its percept. In particular, they implicitly constrain the temporal integration window w and the typical number of neurons K which give rise to the percept. Comparing neural and behavioral sensitivity alone cannot disentangle these two sources of perceptual integration, so the characteristic equations also involve a measure of choice signals, like those assessed by the classic experimental measure of choice probabilities. We then propose a statistical method of analysis which allows to recover the typical scales of integration w and K from finite numbers of recorded neurons and recording trials, and show the efficiency of this method on an artificial encoding network. We also study the statistical method theoretically, and relate its laws of convergence to the underlying structure of neural activity in the population, as described through its singular value decomposition. Altogether, our method provides the first thorough interpretation of feedforward percept formation from a population of sensory neurons. It can readily be applied to experimental recordings in classic sensory decision-making tasks, and hopefully provide new insights into the nature of perceptual integration.

1 Introduction

Most cortical neurons are noisy, or at least appear so to experimenters. When a sensory neuron’s spikes are recorded in response to a well-controlled stimulus, they will show a large variability from trial to trial. This noisiness has been acknowledged from early on, as a nuisance preventing experimenters from easy access to the encoding properties of sensory neurons. But what is the impact of trial-to-trial sensory noise on the organism itself? This question gained renewed interest a few decades ago, with the generalization of experimental setups recording neural activity from awake, behaving animals (Mountcastle et al., 1990; Britten et al., 1992). In these setups, animals are presented with a set of stimuli f and trained to respond differentially to different values of f , thus providing an (indirect) report of their percept of f . As neural activity and animal behavior are simultaneously monitored, it becomes possible to seek a causal link between the two.

In such setups, one particular hypothesis—which we refer to as the “sensory noise” hypothesis—has proven instrumental in linking neural activity and percepts. It postulates that trial-to-trial noise at the level of sensory neurons is the main factor limiting the accuracy of the animal’s perceptual judgements (Werner and Mountcastle, 1965; Talbot et al., 1968). Indeed, signal detection theory provides the adequate tools to estimate such accuracies. Any type of biological response to a stimulus f —say $r(f)$ —can

be associated to a signal-to-noise ratio (SNR), which measures how typical variations in r due to a change of stimulus f (the *signal*) compare to intrinsic variations of r from trial to trial (the *noise*). When $r(f)$ measures the response of a neuron to stimulus f , the resulting SNR is often called the *neurometric* sensitivity for that particular neuron. Alternatively, $r(f)$ may also be the response of the animal itself to stimulus f . The resulting SNR is called the animal’s *psychometric* sensitivity, which quantifies the animal’s ability to discriminate nearby stimulus values f . Reformulated in terms of SNRs, the “sensory noise” hypothesis states that neurometric sensitivity, computed from the population of sensory neurons under survey, is equal to the psychometric sensitivity for the animal in the task.

Applying this idea, neurometric and psychometric sensitivities have often been computed and compared, in various sensory systems and behavioral tasks (see, e.g., Romo and Salinas, 2003; Gold and Shadlen, 2007, for reference). However, it was progressively realized that most of these comparisons bear no simple interpretation, because the neurometric sensitivity is not a fixed quantity: it depends on how information is read out from the neurons. For example, if the various sensory neurons in the population behave independently one from another, then the overall SNR from the population will essentially be the sum of individual SNRs and thus, the experimenter’s estimate of neurometric sensitivity will depend on how many neurons—say K —they included in their analysis. This intuition still holds in realistic populations where neurons are not independent, with the additional complexity that the evolution of neural SNR with K is very influenced by the correlation structure of noise in the population (Shadlen and Newsome, 1998; Abbott and Dayan, 1999; Averbeck et al., 2006).

More subtly, another parameter has a direct influence on estimated neurometric SNRs: the time scale w used to integrate each neuron’s spike train, to describe the neuron’s activity over the trial (Cohen and Newsome, 2009). Indeed, through the central limit theorem, the more neural spikes are integrated into the readout, the more accurate that readout will be. Adding extra neurons through K , or extra spikes for each neuron through w , will thus have the same type of impact on the readout’s overall SNR. In fact, if all neurons from the population are identical, independent Poisson encoders, one can easily show that the readout’s overall SNR scales with \sqrt{wK} , emphasizing the duality between K and w .

As there is no unique way of reading out information from a population of sensory neurons, a question naturally arises: what type of readout does the organism use? For example, how many sensory neurons K , and what typical integration time scale w , provide a relevant description of the animal’s percept formation? The “sensory noise” hypothesis can precisely be used to answer this question: the ‘true’ neuronal readout for the organism must be the one providing the best account of animal behavior. However, the previous K – w discussion clearly shows that comparing neurometric SNR to psychometric SNR is not sufficient to target the true readout: there will be several combinations of K and w leading to the same overall neurometric SNR, while corresponding to very different extraction strategies by the animal. Thus, an additional experimental measure is required to recover the typical scales of integration of the true readout.

Choice signals are a good candidate for this additional measure. In two-alternative tasks, where the animal must report a binary discrimination of stimulus value (say, $f > 0$ or $f < 0$), choice signals are generally computed in the form of *choice probabilities* (CP) (Green and Swets, 1966; Britten et al., 1996). CP is computed for each recorded neuron individually, and quantifies the trial-to-trial correlation between the activity of that neuron and the animal’s ultimate (binary) choice on the trial, all other features being held constant. In particular, since CP is computed across trials with the same stimulus value (generally uninformative, i.e., $f = 0$), the observed correlations cannot reflect the influence of stimulus on neural activity and animal choice. Instead, a significant CP can only result from the process by which the neuron’s activity influences—or is influenced by—the animal’s forming perceptual decision.

It is intuitively clear that CPs reveal something about the way information is extracted from sensory neurons. For example, if the animal’s percept is built from a single neuron, then that neuron will have a very large CP, because its activity on every trial directly predicts the animal’s percept. Instead, if several independent neurons contribute to form the animal’s percept, then they are all expected to have low

CP value, as the activity of each neuron has only a marginal impact on the animal’s decision. However, converting this intuition about choice signals into a quantitative interpretation was long hampered by the fact that, just like neurometric SNR, CP values are largely influenced by the population’s noise covariance structure. For example, a neuron may not be utilized by the animal to form its percept, and yet display significant CP because its activity is correlated with that of another neuron being utilized. As a result, early studies relating CP values to the animal’s perceptual readout only relied on numerical simulations (Shadlen et al., 1996; Cohen and Newsome, 2009), assuming very specific noise correlation structures that weakened the generalizations of their results. Only very recently have Haefner et al. (2013) provided an analytical expression for CP values in the presence of noise correlations (see section 4.2), opening the door to general, quantitative interpretations of choice probabilities.

In this article, we show how the combined information of animal sensitivity (SNR) and choice signals allows to estimate the typical scales of percept formation by the animal, both across neurons (number of neurons involved K) and in time (integration window w). Our results apply in the standard feedforward model of percept formation, and can be derived for any noise covariance structure in the neural population. We first show how the joint covariance structure of neural activities and animal percept leads to a set of characteristic equations for the readout, which implicitly determine the animal’s perceptual readout policy across neurons and time. Then, we show how these characteristic equations can be used in a statistical form, across the ensemble of trials and neurons available to the experimenter, to determine the typical scales K and w of percept formation from the activity of sensory neurons. This approach is mandatory since experimental measurements can only provide statistical samples of the full neural population. Using an artificial neural network to provide sensory encoding, we show that our method can reliably recover the true scales of perceptual integration, without requiring full measurement of the neural population. Thus, our method can readily be applied to real experimental data, and provide new insights into the nature of sensory percept formation.

2 Methods

2.1 Framework and notations

We place ourselves in a general framework, describing a typical perceptual decision-making experiment (Fig. 1). On each trial, a different stimulus f is presented to the animal (Fig. 1a, top), which then takes a decision according to its internal judgement f^* of stimulus value. Our framework assumes that this percept f^* is directly available to the experimenter on each trial. In real experimental setups, the animal’s report is generally more indirect—typically a binary choice based on the unknown percept f^* . We choose the former approach because it applies generically to most perceptual decision-making experiments, whereas the “choice” part is more dependent on each particular setup. We detail later how both approaches can be reconciled through simple models of the animal’s behavior (section 4.2).

Simultaneously, experimenters record neural activities from a large population of sensory neurons, which is assumed to convey the basic information about f used by the animal to take its decision (Fig. 1a, bottom). Typical examples could be area MT in the context of a moving dot discrimination task (e.g., Britten et al., 1992), area MT or V2 in the context of a depth discrimination task (e.g., Uka and DeAngelis, 2003; Nienborg and Cumming, 2009), or area S1 in the context of a tactile discrimination task (e.g., Hernández et al., 2000). We describe the activity of this neural population on every trial as a point process $\mathbf{s}(t) = \{s_i(t)\}_{i=1\dots N_{\text{tot}}}$, where each $s_i(t)$ is the spike train for neuron i , viewed as a series of Dirac pulses. As an important remark, N_{tot} denotes the full population size, a very large and unknown number. It is *not* the number of neurons actually recorded by the experimenter, which is generally much smaller.

For simplicity, we assume a fine discrimination task, where the different stimulus values f presented to the animal display only moderate variations around a central value, say $f = 0$. This substantially

simplifies SNR computations, because the ‘signal’ part of any response $r(f)$ is then summarized by its slope in $f = 0$: $\partial_f E(r|f)|_{f=0}$, where $E()$ denotes the average response over trials. We assume that this linearization with f can be performed both for the psychometric report f^* , and for individual neuron activities. This is mostly a convenience though, and the framework could be generalized to more complex dependencies on stimulus f .

From the raw data of f^* and $s(t)$ on each trial, a number of measures are routinely used to describe neural activity and animal behavior. First, the psychometric sensitivity Z^* describes the animal’s accuracy in distinguishing nearby frequency values. It can be computed from the distribution of (f, f^*) across trials (Fig. 1b), according to the formula:

$$Z^* = \frac{1}{\langle \text{Var}(f^*|f) \rangle_f}, \quad (1)$$

where notation $\langle \cdot \rangle_f$ denotes an average across stimulus conditions. This is exactly the (squared) SNR for random variable $f^*(f)$, assuming that the ‘signal’ term $\partial_f E(f^*|f)$ is equal to 1 because the animal’s average judgement of f is unbiased (the framework easily generalizes to a biased percept).

On the other hand, for each recorded neuron, it is common practice to compute its peri-stimulus time histogram (PSTH) in response to each different tested stimulus (Fig. 1d):

$$\lambda_i(t; f) := E(s_i(t)|f), \quad (2)$$

where E denotes averaging over trials. Since all stimuli f are assumed to be close one from another, the dependency of $\lambda_i(t; f)$ on f is essentially linear, and can be summarized by the (temporal) tuning curve for the neuron (Fig. 1e):

$$\beta_i(t) := \partial_f \lambda_i(t; f). \quad (3)$$

Furthermore, as recent techniques allow the simultaneous recording of many neurons, experimenters also have access to samples from the trial-to-trial covariance structure in the population (Fig. 1c). For every pair of neurons (i, j) and instants in time (t, s) , this covariance structure is assessed through the neurons’ joint peri-stimulus time histogram (JPSTH, Aertsen et al., 1989):

$$\gamma_{ij}(t, s) := \langle \text{Cov}(s_i(t), s_j(s)|f) \rangle_f. \quad (4)$$

We only consider the average covariance structure, over different stimuli f . First, as above, nearby values of f insure that the covariance structure will remain mostly unchanged. Second, trial-to-trial covariances correspond to second-order effects on neural activity, which require several trials to be reliably estimated—another reason to lump data from different stimuli f into a single estimate.

Finally, we can measure a *choice signal* for each neuron, estimating the trial-to-trial covariance of neuron activity $s_i(t)$ with the animal’s choice (Fig. 1f). Since in our framework the animal directly reports its percept f^* , we readily describe the choice signal of each neuron by its *percept covariance* (PCV) curve:

$$\pi_i^*(t) := \langle \text{Cov}(s_i(t), f^*|f) \rangle_f. \quad (5)$$

Again, this covariance information is lumped across the different (nearby) stimulus values f , in order to improve experimental measurement. The PCV curve captures the core intuition behind the more traditional measure of choice probability (CP), while retaining a linear form convenient for analytical treatment. Percept covariance curves are not directly measurable in classic experimental setups where the animal only reports a binary choice ; however their analytical link to available measures such as CPs can be easily derived given simple models of the animal’s decision policy (see section 4.2).

Unlike many characterizations of neural activity that rely only on spike counts, our framework requires an explicit temporal description of neural activity through PSTHs (eq. 2), JPSTHs (eq. 4) and percept covariance curves (eq. 5). Indeed, our method ultimately predict *when*, and *how long*, perceptual

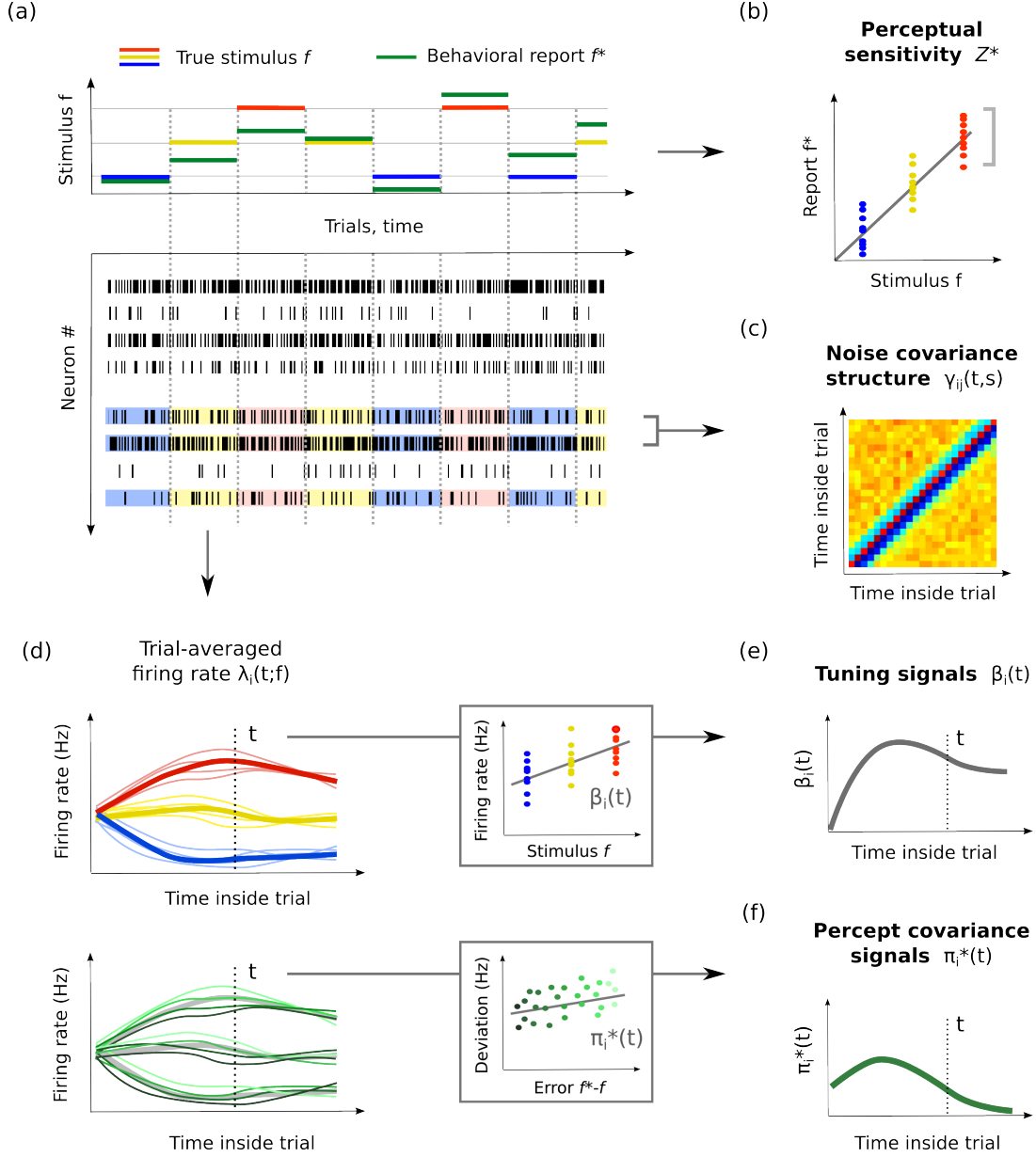


Figure 1. Framework and main experimental measures. (a) Experimental setup. Top: A set of stimulus values f (color-coded as blue, yellow, red) are repeatedly presented to an animal, which reports its percept f^* on each trial (color-coded as green). Bottom: In each session, several task-relevant sensory neurons are recorded simultaneously with behavior. (b) Perceptual sensitivity Z^* is defined as the square SNR of the animal's reports $f^*(f)$. (c) The noise covariance structure can be assessed in each pair of simultaneously recorded neurons, as their joint peri-stimulus histogram (JPSTH). (d) Trial-wise response of a particular neuron. Each thin line is the schematical representation of the spike train on each trial. Segregating trials according to stimulus (top), we access the neuron's peri-stimulus histogram (PSTH) and its tuning curve—shown in panel (e). Segregating trials according to the animal's perceptual error $f^*(f) - f$ (bottom), we access the neuron's percept covariance (PCV) curve—shown in panel (f).

integration takes place in the organism. Readers may feel uncomfortable that the resulting definitions are directly expressed over trains of Dirac pulses. While these notations are fully justified in the framework of point processes (Daley and Vere-Jones, 2007), they describe idealized quantities that cannot be estimated from a finite number of trials, leading to jaggy estimates formed from the collection of Dirac peaks. So in practice, spike trains $s_i(t)$ are computed in temporal bins of finite precision.

2.2 The readout model and its characteristic equations

All experimental measures described above, taken together, provide a full characterization of the joint covariance structure of variables $(\mathbf{s}(t), f^*)$ across stimuli and trials (Fig. 2c). The key argument to exploit these data, which is actually a reformulation of the ‘sensory noise’ hypothesis, is that the animal’s percept f^* is built on every trial from the activity of the sensory neurons, meaning that $f^* = F^*(\mathbf{s})$ for some unknown readout F^* . As a result, each proposed readout F directly yields an estimate for the joint covariance structure of $(\mathbf{s}(t), f^*)$ —through a set of relationships which constitute the readout’s *characteristic equations*. Conversely, since this joint covariance structure is experimentally measurable, it implicitly constrains the nature of the true readout F^* which was applied by the animal. In this section, we introduce a generic form of linear readout, stemming from the standard feedforward model of perceptual integration, and derive its characteristic equations. We show that in theory, these equations totally characterize the readout applied by the animal.

2.2.1 Readout model

We define a generic linear readout from the activity of sensory neurons $\mathbf{s}(t)$ (Fig. 2a), based on a given readout vector: $\mathbf{a} = \{a^i\}_{i=1\dots N_{\text{tot}}}$, a given integration kernel with normalized shape h and time constant w : $h_w(t) := w^{-1}h(\tau/w)$, and a given readout time t_R :

$$\hat{f}(t_R) := \sum_i \int_{u>0} a^i s_i(t_R - u) h_w(u) du. \quad (6)$$

The readout is noted \hat{f} , as it must ultimately be an estimator of stimulus value f . We explicitly note the dependence on t_R to emphasize that $\hat{f}(t_R)$ is built from a sliding temporal average of the spike trains ; so that each instant in time yields a potential readout.

This is a classical form of readout from a neural population, which has often been used previously and described as the ‘standard’ model of perceptual integration (Shadlen et al., 1996; Haefner et al., 2013). The temporal parameters h_w and t_R describe how each neuron’s temporal spike train $s_i(t)$ is integrated into a single number describing the neuron’s activity over the trial: $\bar{s}_i = \int_{u>0} s_i(t_R - u) h_w(u) du$. In turn, the percept is built linearly from the population activity as $\hat{f} = \sum_i a^i \bar{s}_i$ through a specific readout vector, or ‘perceptual policy’, \mathbf{a} .

However, traditional studies generally make ad hoc choices for the various constituents of this readout. Most often, \bar{s}_i simply describes the total spike count for neuron i , which in our model corresponds to choosing a square kernel h , and parameters $w = t_R = T$ describing an integration over the full period T of sensory stimulation. As mentionned in the introduction, there is no reason that this should be a relevant description of sensory integration by the organism: the integration window w has a direct influence on predicted SNRs for the readout, and experiments suggest that animals do not always use the full stimulation period to build their judgement (Luna et al., 2005; Stanford et al., 2010).

Instead, we make no assumption on the nature of w and t_R , and view them as free parameters of the model. Then, the model parameters implicitly characterize the typical scales of perceptual integration by the animal. The number of significantly nonzero entries in \mathbf{a} , say K , defines the number of neurons contributing to the percept. The readout window w characterizes the behavioral scale of temporal integration from the sensory neurons, and time t_R characterizes when during stimulation this integration

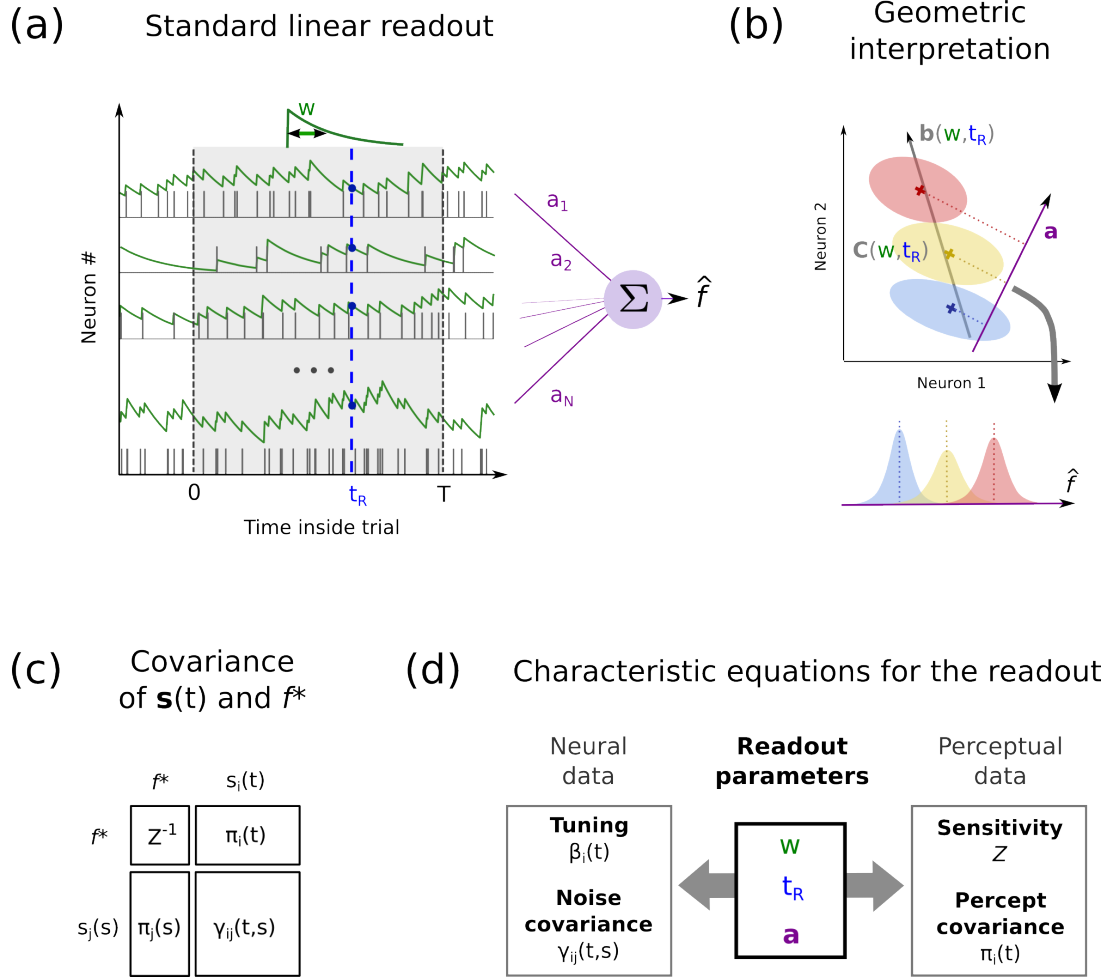


Figure 2. Linear readout and its interpretation. (a) We study a “standard” model of perceptual readout, with two parameters w and t_R defining integration in time, and a readout vector \mathbf{a} defining integration across neurons. (b) Geometric interpretation of the model. The temporal parameters w and t_R define the tuning vector \mathbf{b} and noise covariance matrix \mathbf{C} in the population. Colored ellipses schematize the distribution of neural activities from trial to trial, for the three possible stimulus values. The readout \hat{f} can be viewed as an orthogonal projection of neural activities in the direction given by \mathbf{a} . (c) Sensitivity Z , PCV curves $\pi_i(t)$ and noise covariance JPSTHs $\gamma_{ij}(t, s)$ totally define the joint covariance structure between spike trains $\mathbf{s}(t)$ and percept \hat{f} . (d) Any feedforward readout of neural activities can be viewed as a mapping $\hat{f} = F(\mathbf{s}(t))$, so the true F is implicitly constrained by the covariance data from panel c. In the case of the linear readout model, these constraints are summarized by three characteristic equations, which relate neural and perceptual data through the readout’s parameters w , t_R and \mathbf{a} .

takes place. The exact shape h given to the integration kernel is of less importance ; for conceptual and implementational simplicity we assume it to be a square window. However, we note that (1) other shapes may have a higher biological relevance, such as the decreasing exponential mimicking synaptic integration by downstream neurons, and (2) nothing prevents our method from making h itself a free parameter, provided the data contain enough power to estimate it. Finally, our model can also be extended to versions where extraction time t_R is not fixed, but varies from trial to trial ; this issue is discussed in section 4.3.2.

2.2.2 Characteristic equations for the readout

Thanks to its linear structure, the readout defined in eq. 6 allows for a simple characterization of the covariance structure that it induces between neural activity $\mathbf{s}(t)$ and the resulting percept \hat{f} (Fig. 2b). We show in appendix A that this covariance structure can be summarized by three characteristic equations:

$$1 = \mathbf{b}^\top \mathbf{a}, \quad (7)$$

$$Z^{-1} = \mathbf{a}^\top \mathbf{C} \mathbf{a}, \quad (8)$$

$$\boldsymbol{\pi}(t) = \boldsymbol{\Gamma}(t) \mathbf{a}, \quad (9)$$

where vector \mathbf{b} and matrices $\boldsymbol{\Gamma}(t)$ and \mathbf{C} respectively describe the population's tuning and noise covariance structures, derived from the underlying neural statistics $\boldsymbol{\beta}(t)$ and $\boldsymbol{\gamma}(t)$ introduced in eq. 3-4:

$$b_i(w, t_R) := \int_{u>0} \beta_i(t_R - u) h_w(u) du, \quad (10)$$

$$\Gamma_{ij}(t | w, t_R) := \int_{u>0} \gamma_{ij}(t, t_R - u) h_w(u) du, \quad (11)$$

$$C_{ij}(w, t_R) := \int_{u>0} \Gamma_{ij}(t_R - u) h_w(u) du. \quad (12)$$

We here note the explicit dependency of \mathbf{b} , $\boldsymbol{\Gamma}(t)$ and \mathbf{C} on the temporal parameters of the readout w and t_R . We will generally omit it in the sequel. Thus, the right-hand sides of eq. 7-9 depend only on readout parameters w , t_R , \mathbf{a} and on the statistics of neural activity, independently of the animal's percept.

On the other hand, the left-hand sides of eq. 7-9 describe experimental quantities related to the readout's resulting percept \hat{f} . The first line describes the average tuning of \hat{f} to stimulus f , that is $\partial_f \mathbb{E}(\hat{f} | f)$, which is equal to 1 because we assume that \hat{f} is unbiased. The second line expresses the resulting sensitivity Z for the readout, defined as in eq. 1. It reveals the dual influence of the number of neurons (through \mathbf{a}) and integration window w on the readout's overall sensitivity: indeed, under mild assumptions, the covariance matrix \mathbf{C} scales with w^{-1} (see appendix A.2.1). Finally, the third line expresses the resulting covariance between \hat{f} and the activity of each neuron $s_i(t)$, defined as in eq. 5. This is essentially the relationship already revealed by Haefner et al. (2013), that choice probabilities are related to readout weights through the noise covariance matrix ; however, our formalism focuses on the simpler linear measure of PCV curves, and explicitly takes time into account.

Both the neural measures $\boldsymbol{\beta}(t)$ and $\boldsymbol{\gamma}(t, s)$ on the right-hand side, and the percept-related measures Z and $\boldsymbol{\pi}(t)$ on the left-hand side, can be estimated from data. As a result, the characteristic equations define an implicit constraint on the readout parameters w , t_R and \mathbf{a} (Fig. 2d). Actually, if the readout model in eq. 6 is true, and precise measures are available for all neurons in the population, one sees easily that these constraints would uniquely determine the readout parameters. Indeed, for fixed parameters w and t_R , eq. 7 and 9 impose linear constraints on vector \mathbf{a} . These constraints are generally overcomplete, since \mathbf{a} is N_{tot} -dimensional, while each time t in eq. 9 provides N_{tot} additional linear constraints. Thus, in general, a solution \mathbf{a} will only exist if one has targeted the true parameters w and t_R , and it will then be unique.

2.3 Estimating the scales of sensory integration

In the previous section we have shown that, in the standard linear model of percept formation, the trial-to-trial covariance structure between spike trains $\mathbf{s}(t)$ and the resulting percept \hat{f} leads to a set of characteristic equations which implicitly define the parameters of the perceptual readout, provided the covariance structure has been fully estimated.

Unfortunately, this direct approach makes a fundamental assumption which cannot be reconciled with real, experimental recordings: it assumes we have recorded all neurons from the population under survey, whereas real recordings only ever record from a small subset of that population. Thus we cannot hope to reconstruct the real vector \mathbf{a} , simply because some—probably most—of the neurons contributing to \mathbf{a} were not recorded. Moreover, even across those neurons which were recorded through a series of sessions in a given area, the noise covariance structure can never be fully assessed ; it remains elusive between neurons which were not recorded simultaneously.

For this fundamental reason, the characteristic equations 7-9 should be used with a different perspective than the full recovery of readout parameters. Instead, we propose to exploit the structure of the equations in a statistical approach, with the restricted goal of estimating the typical scales of readout most compatible with recorded data.

2.3.1 Reformulation in terms of neural subensembles

A first necessary step in our approach is to statistically describe the nature of readout vector \mathbf{a} . We are mostly interested in the support of \mathbf{a} , meaning, the number and nature of neurons contributing to percept formation. Thus, we assume that the percept is built only from the activities of an unknown ensemble \mathcal{K} of neurons in the population and that, for given \mathcal{K} and temporal parameters (w, t_R) , the readout vector \mathbf{a} is chosen optimally to maximize the SNR of the resulting percept. Indeed, through this hypothesis, we totally reformulate the problem of characterizing \mathbf{a} in that of characterizing \mathcal{K} ; which allows for much simpler statistical descriptions.

The readout vector \mathbf{a} achieving the maximum sensitivity Z in eq. 8, under the constraints of eq. 7 and having support on \mathcal{K} , is well known from the statistical literature. It is uniquely given by Fisher's linear discriminant formula (Hastie et al., 2009):

$$\mathbf{a}_{\mathcal{K}} = \frac{1}{\mathbf{b}_{\mathcal{K}}^{\top} \mathbf{C}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}}} \mathbf{C}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}}, \quad (13)$$

where $\mathbf{a}_{\mathcal{K}}$, $\mathbf{b}_{\mathcal{K}}$ and $\mathbf{C}_{\mathcal{K}}$ are the versions of vectors \mathbf{a} , \mathbf{b} (eq. 10) and matrix \mathbf{C} (eq. 12) restricted to neuron ensemble \mathcal{K} . By injecting the form (eq. 13) into eq. 8-9 we obtain a new version of the characteristic equations, under the assumption that percept is built optimally from some given ensemble \mathcal{K} , and temporal parameters (w, t_R) :

$$Z(\mathcal{K} | w, t_R) = \mathbf{b}_{\mathcal{K}}^{\top} \mathbf{C}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}}, \quad (14)$$

$$\pi_i(t | \mathcal{K}, w, t_R) = \frac{1}{Z(\mathcal{K})} \mathbf{\Gamma}_{i\mathcal{K}}(t) \mathbf{C}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}}. \quad (15)$$

Z in eq. 14 is the (optimal) sensitivity associated to this particular choice of \mathcal{K} , w and t_R . In eq. 15, $\pi_i(t)$ is the resulting, predicted PCV curve for every neuron i in the population (not only in ensemble \mathcal{K}). $\mathbf{\Gamma}_{i\mathcal{K}}(t)$ is a row vector whose entries are equal to $\Gamma_{ij}(t)$ (eq. 11) for neurons $j \in \mathcal{K}$.

These equations open the door to a statistical description of percept formation in the neural population: we can now parse through a large set of candidate ensembles \mathcal{K} and temporal parameters (w, t_R) , and ask when the predictions for sensitivity (eq. 14) and PCV curves (eq. 15) match their true psychophysical counterparts Z^* (eq. 1) and $\pi_i^*(t)$ (eq. 5). For sensitivity, the straightforward comparison is to require that $Z(\mathcal{K} | w, t_R) \approx Z^*$.

On the other hand, for the PCV equation (eq. 15), it is pointless to search an elementwise match for every neuron i , between the predicted curve $\pi_i(t)$ and its true measure $\pi_i^*(t)$. Indeed, since only a small subset of the neurons have been recorded, no candidate readout ensemble \mathcal{K} will be equal to the true ensemble (say \mathcal{K}^*) that was used by the animal ; and there is no guarantee that the covariance structure between i and \mathcal{K} , which gives rise to prediction (eq. 15), should be similar to that between i and \mathcal{K}^* . Instead, a given set of readout parameters (\mathcal{K}, w, t_R) should be deemed plausible if they predict the correct *distribution* of PCV signals across the population, irrespective of exact neuron identities i . Full distributions are difficult to estimate from finite amounts of data, and we will find the following population *averages* to convey sufficient information:

$$W(t | \mathcal{K}, w, t_R) := E_i \left(b_i(w, t_R) \pi_i(t | \mathcal{K}, w, t_R) \right), \quad (16)$$

$$W^*(t | w, t_R) := E_i \left(b_i(w, t_R) \pi_i^*(t) \right), \quad (17)$$

where E_i denotes averaging over the full population of neurons $i = 1 \dots N_{\text{tot}}$. We will deem a set of readout parameters plausible if they yield $W(\mathcal{K} | w, t_R) \approx W^*(w, t_R)$ ¹. Multiplying each PCV curve by the neuron's tuning b_i (eq. 10) yields more stable estimates for W , as discussed in section 3.2 and appendix B.

2.3.2 Statistical constraints on readout scales

There are many ways to compare the real values of sensitivity and PCV signals, to their predictions given by eq. 14-15. We propose here an ad-hoc method, whose main characteristics are the following: (1) focus mostly on first-order statistics (i.e., means) across the neural population, (2) use arbitrary tolerance values to compare real and predicted data, (3) fit the two indicators sequentially: first SNR, then percept covariance. Due to its simplicity, this method will prove robust to measurements errors arising from finite amounts of data (section 3.3).

Our method is also designed to cope with a fundamental limitation of real recordings: all neurons (ensemble \mathcal{K} , neurons i) contributing to predictions eq. 14-15 must have been recorded simultaneously, to assess their noise covariance structure. This constraint sets a limit on ensemble sizes K which can be easily investigated (but see section 4.4). Moreover, it prevents from estimating the full average of choice signals (eq. 16) predicted by a given ensemble \mathcal{K} —it is only available for simultaneously recorded neurons i . As a result, predictions (eq. 15) from different tested ensembles \mathcal{K} must somehow be aggregated to produce a reliable prediction of choice signals.

We propose that each tested ensemble \mathcal{K} contribute to our estimates in proportion to its ability to account for the animal's sensitivity:

$$P_Z(\mathcal{K} | w, t_R) \sim \exp \left(- \frac{(Z(\mathcal{K} | w, t_R) - Z^*)^2}{2\alpha_Z^2} \right), \quad (18)$$

normalized to insure $\sum_{\mathcal{K}} P_Z(\mathcal{K}) = 1$ across all tested ensembles (w and t_R being fixed). Parameter α_Z is the required tolerance for the fit, set by the experimenter. It is a regularization parameter creating a tradeoff between precision of fit (small α_Z) and reliability of measurements, since a larger α_Z leads to more samples \mathcal{K} with a substantial contribution $P_Z(\mathcal{K})$. When testing our method (section 3) we choose α_Z as 5% of Z^* .

For each tested couple (w, t_R) , we then use $P_Z(\mathcal{K} | w, t_R)$ as a weighting factor over all tested ensembles

¹Note that $W^*(t | w, t_R)$ depends on parameters (w, t_R) only through the neurons' tunings $b_i(w, t_R)$. In practice, as neural activities are rather stationary in time, $W^*(t)$ changes very little for different values of parameters (w, t_R) .

\mathcal{K} , which yields two quantities:

$$\check{K}(w, t_R) := \sum_{\mathcal{K}} P_Z(\mathcal{K} | w, t_R) \text{Card}(\mathcal{K}), \quad (19)$$

$$\check{W}(t | w, t_R) := \sum_{\mathcal{K}} P_Z(\mathcal{K} | w, t_R) E_{i(\mathcal{K})} \left(b_i \pi_i(t | \mathcal{K}, w, t_R) \right), \quad (20)$$

where $E_{i(\mathcal{K})}$ denotes an average across all neurons i available to compute a prediction with eq. 15. These neurons must have been recorded simultaneously to ensemble \mathcal{K} and, in order to produce an unbiased estimate of choice signals in the full population, they should not belong to \mathcal{K} itself.

In eq. 19, $\check{K}(w, t_R)$ is the ensemble size K which most likely explains the animal's sensitivity, given readout parameters (w, t_R) . In eq. 20, $\check{W}(t | w, t_R)$ is the mean prediction for PCV signals $b_i \pi_i(t)$ across neurons i in the population, but stemming only from ensembles \mathcal{K} which are compatible with the animal's sensitivity. Considering quantity $W(t)$ introduced in eq. 16, we see that

$$\check{W}(t | w, t_R) \simeq \sum_{\mathcal{K}} P_Z(\mathcal{K} | w, t_R) W(t | \mathcal{K}, w, t_R). \quad (21)$$

The equality is only approximate, because only neurons i recorded simultaneously to \mathcal{K} are available to estimate $\check{W}(t)$. However, as neurons i are random and we average over many ensembles \mathcal{K} , $\check{W}(t)$ rapidly converges to the quantity described in eq. 21.

Both $\check{W}(t | w, t_R)$ and $W^*(t | w, t_R)$ are temporal signals defined over some interval $[T_{\min}, T_{\max}]$ corresponding to one trial repetition. Defining the L2 norm for such temporal signals as

$$\|x\|^2 = (T_{\max} - T_{\min})^{-1} \int_{t=T_{\min}}^{T_{\max}} x^2(t) dt, \quad (22)$$

we will deem parameters (w, t_R) plausible if they lead to a small value of $\|\check{W}(w, t_R) - W^*(w, t_R)\|^2$. To yield a quantitative estimate of fit, we introduce a tolerance α_W and define the following weighting function:

$$P_W(w, t_R) \sim \exp \left(- \frac{\|\check{W}(w, t_R) - W^*(w, t_R)\|^2}{2\alpha_W^2} \right), \quad (23)$$

normalized to insure $\sum_{w, t_R} P_W(w, t_R) = 1$ across all tested temporal parameters (w, t_R) . Again, tolerance α_W is set arbitrarily by the experimenter. When testing our method (section 3) we choose α_W as 5% of $\|W^*(w, t_R)\|$.

Overall, the statistical method introduced above reduces readout parameters to three numbers: the temporal extraction parameters w and t_R , and the typical number of neurons K used by the readout. Thus, we can now apply a 'brute-force' approach: test all possible combinations (K, w, t_R) , compute the population statistics from eq. 18-23, and target the parameters that provide the best fit. In the next section, we show the validity of this statistical approach, which allows us to recover the typical scales (K, w, t_R) of perceptual integration in an artificial network simulation. We further detail how this statistical approach can be adapted to counteract measurement errors which typically arise in real experiments from the finite number of available trials.

3 Results

3.1 Artificial neural network

In this section, we show how the statistical analysis of sensitivity and choice signals described above allows to recover the scales of integration of the neural readout. Naturally, to assess the validity of our

method, it is necessary to know the true nature of this readout. This can only be achieved through an artificial simulation of sensory integration, where we have full control on neural activities and readout procedure.

We thus implemented an artificial neural network, that encodes some input stimulus f in the spiking activity of its neurons (Fig. 3a). Precise parameters of this network are provided as Supplementary Material (section S1). Briefly, on each trial, 100 input Poisson neurons fire with rate f , taking one of three possible values 25, 30 and 35 Hz. The encoding population *per se* consists of 500 leaky integrate-and-fire (LIF) neurons. 100 of these neurons receive sparse excitatory projections from the input Poisson neurons, which naturally endows them with a positive tuning to stimulus f . 100 other neurons receive sparse inhibitory projections from the Poisson neurons, which naturally endows them with negative tuning. The remaining 300 neurons receive no direct projections from the input. Instead, all neurons in the encoding population are coupled through a sparse connectivity with random delays up to 5 ms. Synaptic weights are random and balanced, tuned to ensure overall firing rates around 30 Hz. We implemented and simulated the network using Brian, a spiking neural network simulator in Python (Goodman and Brette, 2008). The statistics of activity for the resulting population are depicted in Fig. 3b,c,e.

We then define the true perceptual readout from this network. We pick a random set of $K^* = 40$ neurons in the population, whose activity is integrated over $w^* = 50$ ms and read out at time $t_R^* = 80$ ms, on each stimulus presentation (each presentation lasting 500 ms). The resulting estimator f^* of stimulus value is built optimally given these constraints, through Fisher linear discriminant analysis (eq. 13)². This leads to a ‘psychometric’ sensitivity $Z^* \approx 0.06 \text{ Hz}^{-2}$, meaning that the network can typically discriminate variations of $(Z^*)^{-1/2} \approx 4.2 \text{ Hz}$ in the input f . (For comparison, over the same integration period w^* , the 100 input Poisson neurons can discriminate variations around 2.5 Hz.) We then compute a PCV for every recorded neuron, measuring its trial-to-trial covariance with estimator f^* (Fig. 3d). Applying the statistical method described above, our goal is now to recover the scales (K^*, w^*, t_R^*) of perceptual integration, on the basis of the experimental measures depicted in (Fig. 3c-e).

3.2 Validation on full population data

To show the theoretical validity of our analysis, we first apply it to a situation where the trial-to-trial covariance of neurons and percept f^* (eq. 1-5) has been fully measured, with high precision³. In particular, we assume full knowledge of the noise covariance structure $\gamma(t, s)$ in our population (eq. 4). Actually, in these unrealistic conditions, the statistical analysis described above is not useful: instead, one could directly solve the characteristic equations (eq. 7-9) to recover the readout parameters \mathbf{a} , w and t_R . However, this is a necessary first step to verify that our method is not flawed theoretically. Do the statistical quantities introduced in eq. 18-23 allow to recover the true scales of integration (K^*, w^*, t_R^*) ?

Assuming a square integration kernel h , we test a set of candidate temporal integration windows w from 10 to 100 msec, and a set of candidate readout times t_R from 10 to 200 msec, all in steps of 10 msec. We then pick randomly candidate neural ensembles \mathcal{K} , of sizes ranging from 2 to 90 neurons, with 50 different random ensembles for each tested size K . For each tested parameters (w, t_R) , we compute the distribution of predicted SNRs $Z(\mathcal{K})$ given by eq. 14, across all candidate neural ensembles (Fig. 4a). Each neural sample \mathcal{K} is then associated to a weight $P_Z(\mathcal{K})$ describing the goodness of fit to the true SNR Z^* (eq. 18). Following eq. 19-20, this yields an estimate for the best-fitting population size $\check{K}(w, t_R)$ (Fig. 4a, dashed vertical line) and mean PCV curve $\check{W}(t | w, t_R)$ (Fig. 4b). Since we assume full knowledge of experimental data, all 500 neurons i are involved in estimating $\check{W}(t)$, independently of ensemble \mathcal{K} .

In Fig. 4c, we show the estimated population size $\check{K}(w, t_R)$ as a function of w and t_R . It shows the mark of the K - w tradeoff on sensitivity, mentioned in the introduction: smaller integration windows w

²To avoid overfitting issues, the trials used to learn the optimal f^* are not used in the subsequent analysis.

³To compute these estimates, all 500 neurons were simultaneously monitored over 16,500 repetitions of each stimulus condition (not counting the trials used to train estimator f^*). Using bootstrap resamplings over stimulus repetitions, we checked that the resulting measures were virtually error-free.

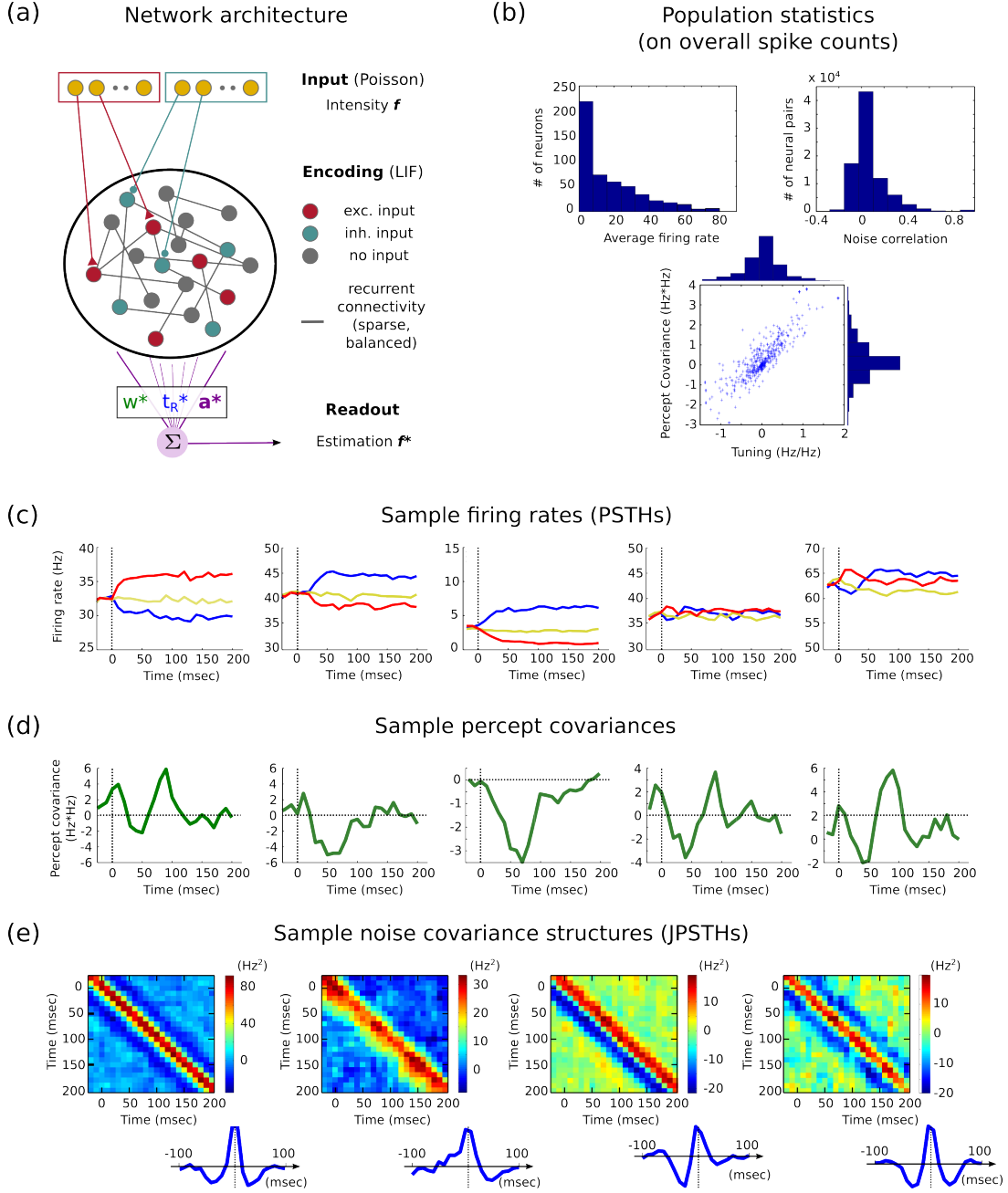


Figure 3. Artificial neural network used for testing the method. (a) Network architecture. The encoding layer consists of LIF neurons coupled through a sparse, balanced recurrent connectivity with random delays. Stimulus f is the firing intensity of a group of input Poisson neurons, which project sparsely into two subpopulations of the encoding layer (neurons excited by the input vs. neurons inhibited by the input). The majority of neurons in the encoding layer receive no direct projection, but can still acquire stimulus tuning through the recurrent connections. A “true” readout f^* is produced on every trial on the basis of true parameters w^* , t_R^* and K^* —which should be retrieved by our method. (b) Classic population statistics in the encoding layer, for neural spike counts over a trial. (c) Sample PSTHs from the encoding layer. Model neurons display varied firing rates, and tunings of different polarities. (d) Sample PCV curves for the same neurons as panel c, computed by correlating each neuron’s spikes with the true readout f^* . (e) Sample JPSTHs (noise correlations) for pairs of neurons in the encoding layer. Inset: corresponding cross-correlograms, obtained by projection along the diagonal.

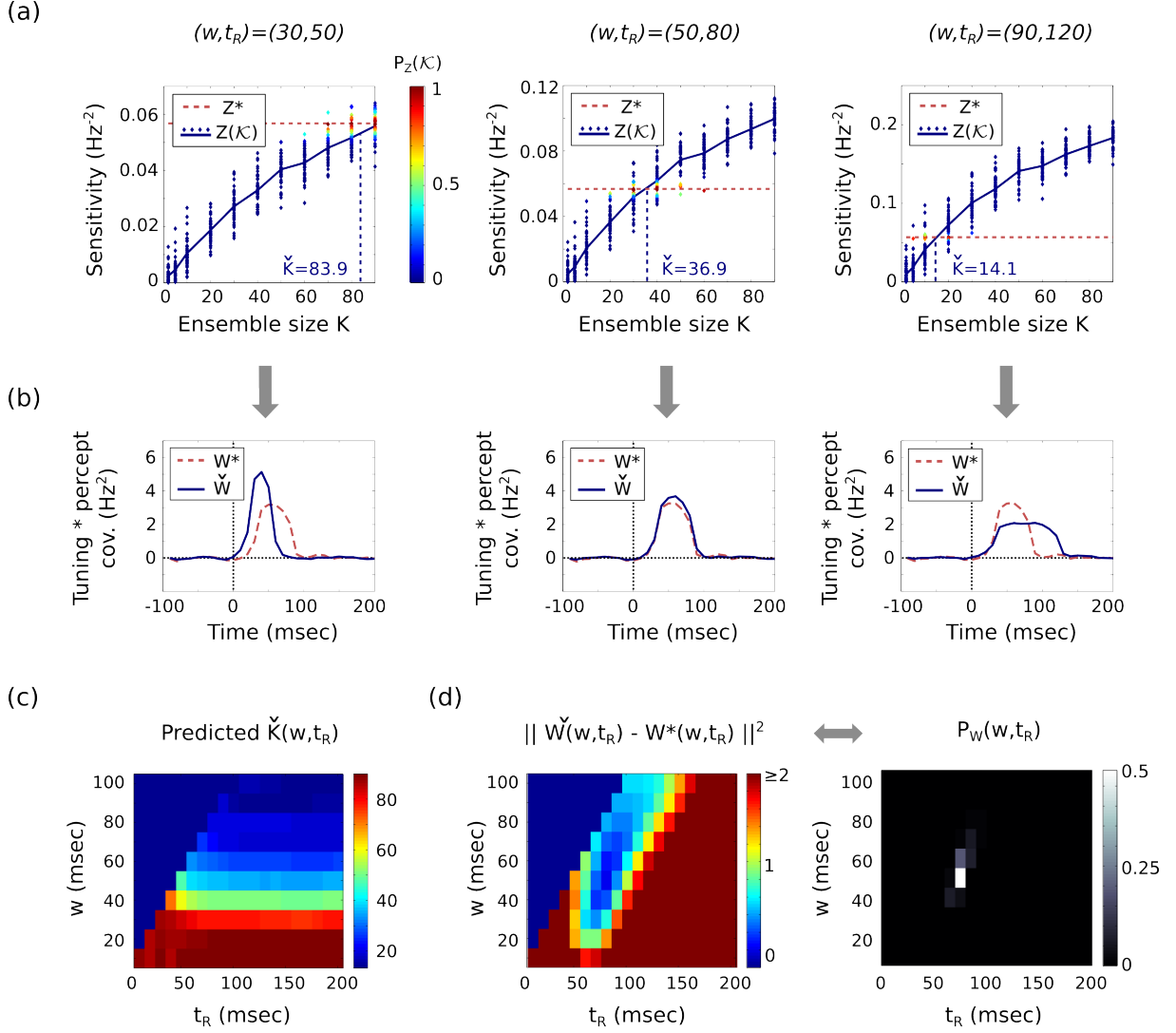


Figure 4. Statistical recovery of readout parameters: noiseless measures. (a) For each tested temporal parameters (w, t_R) , predicted sensitivities $Z(K)$ are computed for several candidate readout ensembles K of varying sizes. The goodness of fit to true sensitivity Z^* defines a weighting function $P_Z(K)$ across ensembles. (b) The weighting function is used to compute a compound prediction $\tilde{W}(t)$ for the average PCV signal in the population, which is compared to the true average $W^*(t)$. The three columns in panels a-b correspond to different candidates (w, t_R) for temporal integration. (c) Best-fitting ensemble size \tilde{K} depending on candidate parameters (w, t_R) . The $K - w$ tradeoff on sensitivity is clearly visible. (d) Goodness of fit of PCV signals depending on candidate parameters (w, t_R) shows a clear optimum around the true parameters of the readout. (e) Same as panel d, but transformed into a weighting function $P_W(w, t_R)$ over candidate temporal parameters.

require larger ensemble sizes K to account for the animal’s sensitivity. In Fig. 4d, we show two measures of the resulting fit between $\check{W}(t | w, t_R)$ and its true value $W^*(t | w, t_R)$. In the first panel, we plot the plain L2 norm between the two temporal signals (using $T_{\min} = -100$ msec and $T_{\max} = 200$ msec as integration bounds). In the second panel, we reexpress this L2 norm as a weighting $P_W(w, t_R)$ over the set of tested temporal parameters (eq. 23). Applying this final weighting over candidate values w , t_R , and $\check{K}(w, t_R)$ yields numerical estimates for the scales of the readout:

$$\begin{aligned}\hat{w} &= 54 \pm 9 \text{ msec} \\ \hat{t}_R &= 81 \pm 6 \text{ msec} \\ \hat{K} &= 34.7 \pm 8.7\end{aligned}$$

These estimates are very close to the true values w^* , t_R^* and K^* , showing the theoretical validity of this approach. The estimated \hat{K} is somewhat smaller than its true value $K^* = 40$, however this is no bias in our method: it simply means that the 40 neurons chosen randomly as the source of percept were slightly less sensitive than the ‘average’ 40 neurons in the population.

We also remind that these estimates depend on the tolerance levels fixed by the experimenter to compute $P_Z(\mathcal{K})$ (eq. 18) and $P_W(w, t_R)$ (eq. 23). Numerically, we find the resulting mean estimates to be rather stable across a range of sensible tolerances. On the other hand, the resulting error bars—which are obtained as second-order moment of the quantities weighted by $P_W(w, t_R)$ —only describe the typical variations of the parameters that lead to estimates within the fixed tolerances. In particular, driving the tolerances to zero always drives the error bars to zero, even though the predicted averages may become false as too little data enter their computation.

Why does the method work? Essentially, it proceeds in two successive steps. First, (w, t_R) being held fixed, it uses SNR information to target plausible neural ensembles \mathcal{K} (Fig. 4a,c). Since the readout is assumed to be optimal, the mean SNR can only increase with the size K of the ensembles considered (Fig. 4a, plain blue curve). Plausible ensembles \mathcal{K} are those lying near the crossing of this curve with the true ‘psychometric’ SNR (Fig. 4a, dashed red curve). For a straightforward application of our method, this crossing should occur within the typical ensemble sizes K tested—which are, in practice, limited by the number of simultaneously recorded neurons. In section 4.4, we discuss possible extensions of the method to the case where the crossing does not occur.

Second, (w, t_R) being still fixed, an average PCV prediction $\check{W}(t | w, t_R)$ is built, using the neural ensembles \mathcal{K} targeted above, and compared to the true mean PCV curve $W^*(t | w, t_R)$. It is not trivial that this comparison should work. To simplify the argumentation, let us assume that parameters w and t_R are fixed at their true values w^* and t_R^* . On the one hand, since the true percept is built from some (unknown) neural ensemble \mathcal{K}^* , we have $W^*(t) = W(t | \mathcal{K}^*)$, using the notations of eq. 16-17. On the other hand, the prediction $\check{W}(t)$ is built as a compound mean of $W(t | \mathcal{K})$ over several candidate ensembles \mathcal{K} (see eq. 21). As our method requires a match between $W^*(t)$ and $\check{W}(t)$, it implicitly supposes that all ensembles \mathcal{K} contributing to $\check{W}(t)$ lead to very similar population averages $W(t | \mathcal{K})$.

Predicted curves $W(t | \mathcal{K})$ are generally not available experimentally. However, we can compute them in our full-data simulation (Fig. 5). We find that, amongst ensembles \mathcal{K} of similar size K , the i -population means of $b_i \pi_i(t | \mathcal{K})$ rapidly converge to a single curve, independently of ensemble \mathcal{K} (Fig. 5b). Furthermore, this result is not trivial: when the same analysis is performed on the plain PCV curves, not multiplied by tuning b_i , the convergence does not occur anymore (Fig. 5a), or at least not as fast. In Fig. 5c, we plot the ratio between the variance of curves accross different ensembles \mathcal{K} , and the power of the mean curve, accross all ensembles \mathcal{K} of same size K . This ratio quickly drops to zero for the tuning-multiplied PCV curves (blue), but not for the plain PCV curves (green).

To summarize, it is crucial for our method that each PCV curve $\pi_i(t)$ be multiplied by the neuron’s tuning b_i before computing population averages. Aside from the experimental observations of Fig. 5,

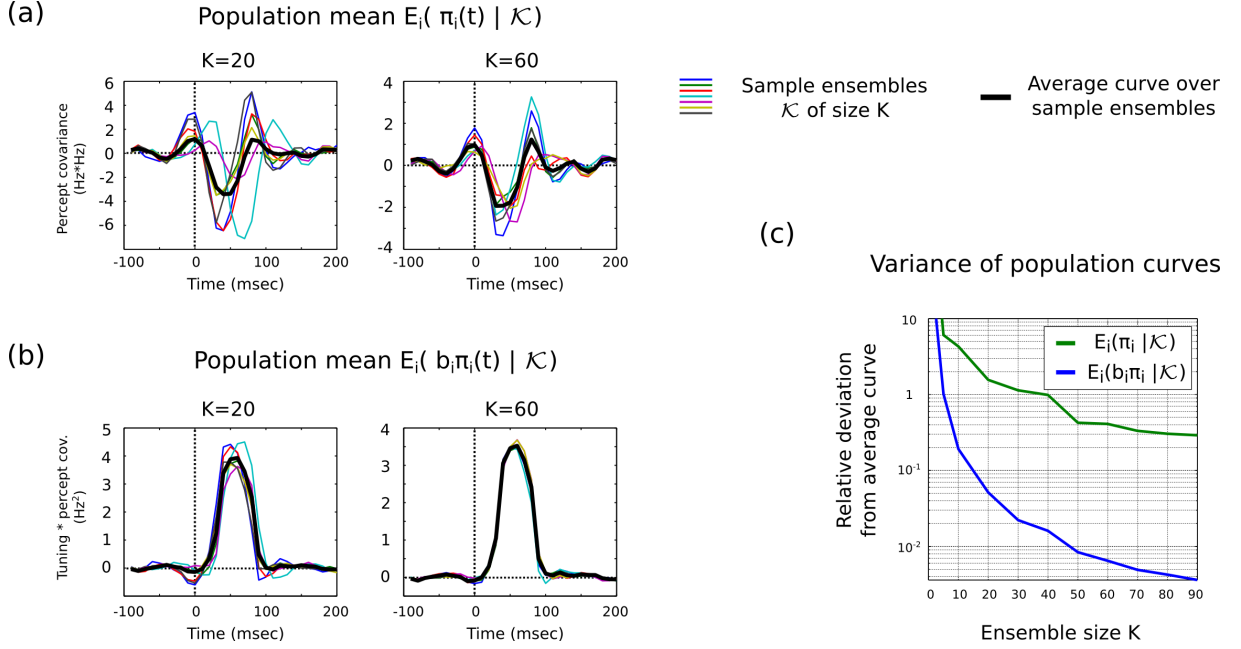


Figure 5. Mean percept covariance curves depend on readout ensemble \mathcal{K} . (a) The mean value of PCV curves $\pi_i(t)$ across neurons i in the population depends strongly on the readout ensemble \mathcal{K} giving rise to the percept. (b) The mean value of tuning-multiplied PCV curves $b_i \pi_i(t)$ depends much less on the exact ensemble \mathcal{K} , only on its size. This justifies our definition for the mean PCV curve $W(t|\mathcal{K})$ (eq. 16). (c) Relative variance of mean PCV curves, across readout ensembles \mathcal{K} of the same size. It is defined as the average of $\|W(\mathcal{K}_1) - W(\mathcal{K}_2)\|^2$ across all ensembles $(\mathcal{K}_1, \mathcal{K}_2)$ of similar size, divided by $\|E_K W(\mathcal{K})\|^2$, power of the average curve across ensembles of size K . For the tuning-multiplied version of $W(t|\mathcal{K})$ (blue), this ratio quickly drops to zero. This is not the case for the plain mean $E_i(\pi_i(t)|\mathcal{K})$ (green).

several arguments justify this operation. First, it is well-known experimentally that choice signals and tuning for individual neurons are often positively correlated at the population level (Britten et al., 1996; Uka and DeAngelis, 2003). Intuitively, this is because positively-tuned neurons contribute positively to stimulus estimation, and conversely for negatively-tuned neurons. The strong population-wide correlation is indeed present in our simulated network (Fig. 3b). As a result, the population average for $b_i\pi_i(t)$ is expected to be mostly positive (Fig. 5b), which diminishes possible variations from one ensemble \mathcal{K} to the other. Second, theoretical arguments (appendix B and Supplementary Material S2) show that $b_i\pi_i(t)$ is a form better suited to compute an i -population average. It can be shown to be positive under mild assumptions, and its laws of convergence can be related to the overall spectrum of covariance in the population.

3.3 Validation on finite data

Having shown the theoretical efficiency of the statistical quantities introduced above in retrieving the correct scales of perceptual integration, we now test our method on its real purpose: recovering the scales from incomplete experimental data (Fig. 6). We thus limit our measures to 150 repetitions for each tested stimulus. Furthermore, we split our population in 5 ensembles of 100 ‘simultaneously recorded’ neurons, so that noise covariance information (eq. 4) is only available between neurons belonging to the same ensemble. We use the same candidate values for parameters w , t_R and K as before, picking 50 candidate ensembles \mathcal{K} for each tested size K . Neurons in \mathcal{K} always belong to the same ‘simultaneous ensemble’, which is picked randomly. Finally, for each ensemble \mathcal{K} , we consider 10 additional neurons i , from the same ‘simultaneous ensemble’ but segregated from neurons in \mathcal{K} , to compute the PCV prediction $\check{W}(t)$ (eq. 20).

The method then proceeds as above, save a couple of modifications due to the incompleteness of the data. First, concerning SNR computations (eq. 14), the estimated covariance matrix $C_{\mathcal{K}}$ may turn out to be rank-deficient up to numerical precision (although it should be full-rank theoretically, since the number of trials (450) is larger than the largest tested size K). We thus replace its inverse $C_{\mathcal{K}}^{-1}$ by its Moore-Penrose pseudo-inverse, with the default numerical tolerance of our mathematical software (Matlab). Even so, we observe a global overestimation of predicted sensitivities Z , compared to their values in the full-data case (dashed blue lines in Fig. 6a, reproduced from Fig. 4a). This overfitting is a well-known feature when estimating Fisher sensitivity from insufficient data (Raudys and Duin, 1998; Hoyle, 2011).

Second, concerning mean PCV predictions (eq. 20), our final estimates $\check{W}(t)$ become noisy, reflecting the jaggedness of the underlying neural measures due to insufficient trials (Fig. 6b). This jaggedness is problematic, as it artificially increases measured values for the divergence $\|\check{W}(w, t_R) - W^*(w, t_R)\|^2$, which is our final criterion to retrieve plausible values of (w, t_R) . However, this effect can be largely compensated by resorting to resampling over trials (bootstrap). More precisely, for each tested parameters (w, t_R) , we may describe our noisy measures in the form:

$$\begin{aligned}\check{W}^{(meas)}(t) &= \check{W}^{(real)}(t) + \eta(t), \\ W^{\star, (meas)}(t) &= W^{\star, (real)}(t) + \eta^*(t),\end{aligned}$$

where $\eta(t)$ and $\eta^*(t)$ describe our (unknown) measurement errors on \check{W} and W^* . From this follows the estimate:

$$\mathbb{E}\|\check{W}^{(meas)} - W^{\star, (meas)}\|^2 = \|\check{W}^{(real)} - W^{\star, (real)}\|^2 + \mathbb{E}\|\eta\|^2 + \mathbb{E}\|\eta^*\|^2, \quad (24)$$

where \mathbb{E} denotes the (theoretical) expectancy over the set of trials giving rise to measures \check{W} and W^* . This estimate is based on the assumption that measurement errors $\eta(t)$ and $\eta^*(t)$ are independent, which is likely to be the case given that $\check{W}(t)$ stems from predictions (on the basis of neural tunings and noise covariances) whereas $W^*(t)$ stems from measurements of the true PCV curves. All terms involving an

expectancy E in eq. 24 can be estimated by resampling with replacement over the set of recorded trials. By computing their difference, we thus get a corrected estimate for $\|\check{W}^{(real)} - W^{*,(real)}\|^2$. This is the estimate plotted in Fig. 6d. A drawback of this method is that the resulting estimate may become (slightly) negative when the underlying match between \check{W} and W^* is “too good”. However, this does not prevent from estimating the resulting weighting function $P_W(w, t_R)$ (eq. 23), accepting that some terms in the exponential may become (slightly) positive.

The final results, in Fig. 6, show that our method is still able to recover the most plausible scales of perceptual integration. Using the same tolerances as previously (5% of the power of the true measures), we find the following final estimates:

$$\begin{aligned}\hat{w} &= 50 \pm 8 \text{ msec}, \\ \hat{t}_R &= 81 \pm 6 \text{ msec}, \\ \hat{K} &= 28.3 \pm 5.2,\end{aligned}$$

again very close to the true scales of the readout. Notably, \hat{K} is smaller than its prediction in the full-data analysis: this is a consequence of the slight overfitting on estimated SNRs, which leads to an underestimation of the population size K required to match the psychometric SNR. The issue remains minor in this setup ; however we note that standard cross-validation and regularization techniques exist, that respectively assess and counteract the effects of overfitting (Hastie et al., 2009).

In conclusion, the statistical method introduced above allows to overcome the missing data inherent to realistic recordings, by integrating information from all recorded neurons into a few reliable statistical estimators.

4 Discussion

4.1 Link with previous literature

We have proposed a framework to interpret sensitivity and choice signals in a standard model of perceptual decision-making. The purpose of our study is to help understand how perceptual integration takes place from a full sensory neural population. This question requires, not only to compute neurometric sensitivities or choice signals for individual neurons, but also to integrate these measures in a single big picture of how information is read out from the population as a whole.

The sensitivity to stimulus achievable by a neural population has received much attention, both experimental and theoretical. It was progressively realized that (1) the structure of noise correlations influences the amount of information that can be extracted from a neural population, and (2) the linear readout maximizing sensitivity is generally not a simple average of neural activities, but rather an adequate weighting optimizing the ratio between signal and noise extracted from the population, which corresponds to Fisher’s linear discriminant (see Abbott and Dayan, 1999; Averbeck et al., 2006, and references therein). Similarly, the role of time window w used to integrate the spike counts of each neuron has long been acknowledged to have a direct effect on the overall estimated sensitivity (see, e.g., Britten et al., 1992; Uka and DeAngelis, 2003; Cohen and Newsome, 2009; Price and Born, 2010).

Choice signals have also received much attention since their first measurements, in the form of choice probabilities (Britten et al., 1996). The temporal evolution of choice signals is routinely computed to qualitatively establish the instants in time when a given population covaries with the animal’s percept (de Lafuente and Romo, 2006; Price and Born, 2010). Recently, the specific temporal evolution of CP signals during a depth discrimination task has cast doubt on the traditional, feedforward interpretation of CP signals (Nienborg and Cumming, 2009, see section 4.3). However, very little studies have *quantitatively* interpreted CP signals so far, because no analytical relationship was available to interpret their values.

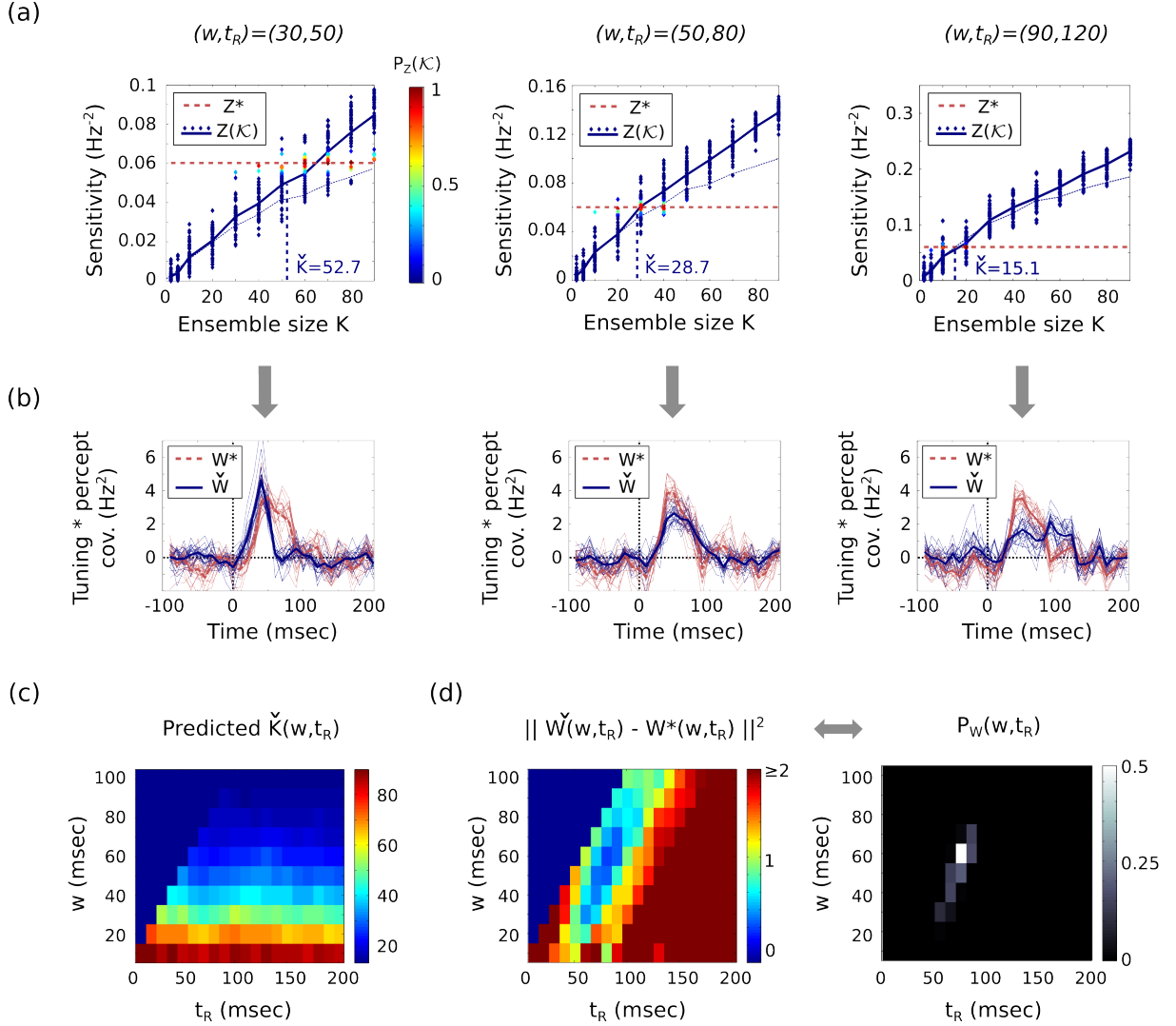


Figure 6. Statistical recovery of readout parameters: noisy measures. Same legends as Fig. 4, but with modifications specific to small sample data. In panel b, the thin curves are different versions obtained through bootstrap resampling over trials, and the thick curve is the average across bootstrap samples. In panel d, the L2 norm is corrected for measurement errors, using the bootstrap samples and eq. 24. With this modification, our method can recover the true readout parameters on the basis of finite amounts of data.

Only recently have Haefner et al. (2013) derived the analytical expression of CPs in the standard model of perceptual integration (see section 4.2).

To the best of our knowledge, only one study has explicitly proposed to jointly use sensitivity and choice signals, as two independent constraints characterizing the underlying neural code. In this seminal study, Shadlen et al. (1996) proposed a feed-forward model of perceptual integration in visual area MT responding to a moving dots stimulus, and studied how the population’s sensitivity and individual neuron CPs vary as a function of model parameters such as the number of neurons, strength of noise correlations, etc. In section 2.2, we have formalized this intuition of Shadlen et al. (1996), by showing that sensitivity and choice signals are two distinct, constitutive elements of the joint covariance structure between percept \hat{f} and neural activity $\mathbf{s}(t)$ (Fig. 2c). The third constitutive element is the noise covariance structure of $\mathbf{s}(t)$ itself, a result also intuited by Shadlen et al. (1996) even though they assumed an oversimplified, homogeneous noise correlation matrix.

Unlike most previous theoretical studies on the subject, we explicitly modeled all neural activities in time. Indeed, this is the only way of targeting the instants of sensory stimulation which contribute to percept formation, and thus to decipher to K - w tradeoff on sensitivity. Finally, the statistical approach developed in section 2.3 is, to our knowledge, the first attempt to build inhomogeneous, partial measures of neural activity into a quantitative interpretation of percept formation from the full neural population.

4.2 Choice signals in realistic experiments

Our model, as presented above, assumes a direct perceptual report of stimulus value f^* on every trial. Real experiments generally involve a more indirect report: to allow easier task learning by the animal, the report is always binary. In the classic random dot motion discrimination task (Britten et al., 1992), a monkey is visually presented with a set of randomly moving dots whose overall motion is slightly biased towards the left ($f < 0$ in our notations) or towards the right ($f > 0$). The monkey must then press either of two buttons depending on its judgement of the overall movement direction. In another classic task (Mountcastle et al., 1990), monkeys must discriminate the frequencies f_1 and f_2 of two successive vibrating stimuli on their fingertip. They must press one button if they consider that $f_1 > f_2$, and the other button otherwise.

Thus, classic choice signals such as CP only measure the covariation between the spike train of each neuron and the animal’s binary choice c on each trial. To infer anything about the animal’s underlying *percept* f^* , it is also necessary to assume a behavioral model describing how the monkey takes a binary decision, on every trial, on the basis of its sensory percept. Most often, this behavioral model is implicitly assumed to be optimal. For example, in the random dot motion task, it is generally assumed that $c = H(f^*)$ (Heavyside function), which is clearly the optimal policy if the animal has no further information about f . In the two-frequency task, the optimal behavioral model would be $c = H(f_1^* - f_2^*)$. However, in the real experiment, the monkeys have to memorize f_1 for a few seconds before f_2 is presented, so potential effects of memory loss may also come into play. More generally, behaving animals can display biases, lapses of attention, various exploratory and reward-maximization policies that lead to deviations from the optimal behavioral model. To summarize, choosing a relevant behavioral model is a connex problem that cannot be addressed here, and that will vary depending on the task and individual considered.

However, for most tractable behavioral models, the predicted sensitivities and choice signals will ultimately rely on the quantities introduced in this article. To take the simplest example, we focus on the random dot motion task with optimal policy $c = H(f^*)$ —as assumed in most models of the task—and make the classic assumption that the statistics of f^* (given f) are Gaussian (Fig. 7a). This model predicts the following psychometric curve (probability of button presses as a function of stimulus value):

$$\mathbb{E}(c|f) = \mathbb{P}(f^* > 0|f) = \Phi(\sqrt{Z^*}f),$$

where Φ is the standard cumulative normal distribution, and Z^* is the square SNR for f^* , as defined in

eq. 1. Thus, Z^* used in our model can easily be retrieved from experimental measures of the psychometric curve.

Same results hold for choice signals. Generally, choice signals are directly computed over some temporal average \bar{s}_i of the underlying spike trains. Choice probability for every neuron i measures the area under the ROC curve between the two distributions of \bar{s}_i , respectively conditioned on $c = 0$ and $c = 1$ (Green and Swets, 1966). Recently Haefner et al. (2013) have shown that, assuming (1) multivariate Gaussian statistics between \bar{s} and f^* , and (2) the optimal behavioral model $c = H(f^*)$, choice probability can be analytically expressed as:

$$\text{CP}_i \simeq \frac{1}{2} + \frac{\sqrt{2}}{\pi} \frac{\sqrt{Z^*} \pi_i^*}{\sigma(\bar{s}_i)},$$

a formula virtually exact over the full range of plausible CP values. The rightmost fraction is nothing but the Pearson correlation between variables \bar{s}_i and f^* . The numerator involves the linear covariance between \bar{s}_i and f^* which is, in our notations, the temporally averaged PCV curve π_i^* . The authors further derived that, in the standard model of percept formation with readout vector \mathbf{a} , this term is given by $\bar{\pi}^* = \mathbf{Ca}$, which is exactly the PCV characteristic equation (eq. 9) in its temporally-averaged form. The CP formula involves a normalization by $\sigma(\bar{s}_i)$, the standard deviation of spike count \bar{s}_i . This prevents from a straightforward extension of the formula in time, because $\sigma(\bar{s}_i)$ tends to infinity as the integration window used to compute \bar{s}_i tends to zero.

A simpler measure of choice signals is the choice-conditioned difference in firing rate (Britten et al., 1996), which can be computed for every individual neuron i as $\Delta_i(t) := E(s_i(t) | c = 1) - E(s_i(t) | c = 0)$. Under the same assumptions as above (Gaussian statistics for $\mathbf{s}(t)$ and f^* , optimal behavioral model), this difference can be analytically expressed⁴ as:

$$\Delta_i(t) = \sqrt{\frac{2}{\pi}} \sqrt{Z^*} \pi_i^*(t). \quad (25)$$

This is very close to the CP formula, but without the additional normalization by $\sigma(\bar{s}_i)$. Thus it directly allows for a simple generalization to temporal signals. Since $\Delta_i(t)$ is easily computable from experimental data, it provides the easiest way of accessing the underlying PCV curves $\pi_i^*(t)$ used in our article.

4.3 Model hypothesis

4.3.1 Linear integration

The readout model (eq. 6) used to analyze sensitivity and choice signals is an instalment of the ‘standard’, feedforward model of percept formation. As such it makes a number of hypothesis which should be understood when applying our methods to real experimental data. First, it assumes that the percept f^* is built linearly from the activities of the neurons. There is no guarantee that this is the case during real percept formation, but linearity is an unavoidable ingredient to make quantitative predictions at the population level. Even if the real percept formation departs from linearity, fitting a linear model will most likely retain meaningful estimates for the coarse information (temporal scales, number of neurons involved) that we seek to estimate in this work.

More precisely, the model in eq. 6 assumes that spikes are integrated using a kernel separable across neurons and time, that is $A^i(t) = a^i h_w(t)$. Theory does not prevent from studying a more general integration, where each neuron i contributes with a different time course $A^i(t)$. The readout’s characteristic equations are derived equally well in that case. Rather, assuming a separable form reflects (1) the intuition that the temporal components of integration are rather uniform across the population, and (2)

⁴Relying on the general formula $E(X_1 | X_2 > 0) = \sqrt{2\pi^{-1}}\rho$, applicable to any bivariate normal variables (X_1, X_2) with means 0, unitary variances, and correlation coefficient ρ . We note that the assumption of normality is violated at small time scales because $s_i(t)$ is clearly not Gaussian in that case. However, in practice, $\Delta_i(t)$ is always computed with a minimal amount of temporal smoothing which resolves this potential issue.

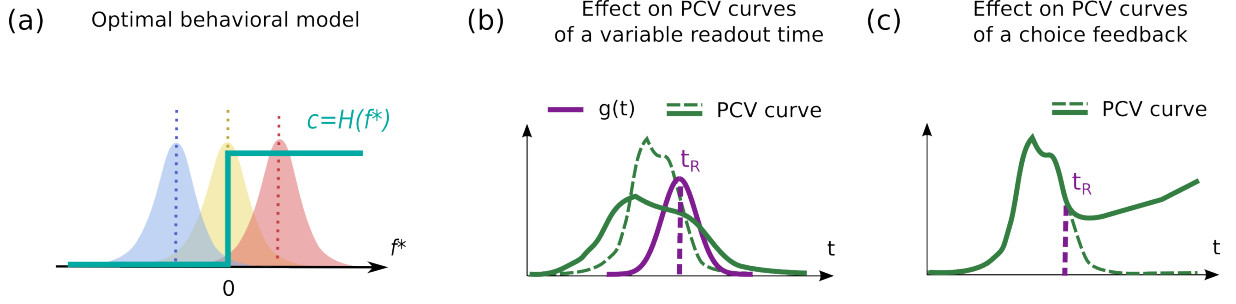


Figure 7. Discussion. (a) Classic behavioral model. If the task is to judge whether $f > 0$, the optimal behavioral policy consists of the simple threshold rule $c = H(f^*)$ (Heavyside function). Furthermore, the trial-to-trial distribution of percept f^* given f (distributions with different colors) is generally assumed to be Gaussian. Under these hypotheses, sensitivity and PCV signals used in this article are directly computed from real experimental data (neurometric curve and choice signals). (b) If readout time t_R varies strongly from trial to trial (with density $g(t)$), it leads to a flattening of PCV signals (thick green curve) compared to the case with deterministic t_R (dashed green curve). (c) If a decision-related signal feedbacks into sensory areas, it leads to a divergence of PCV signals (thick green curve) after the readout time t_R , compared to the case without feedback (dashed green curve).

the impossibility to fit a model with general kernel $A^i(t)$. Instead, we summarize temporal integration from the population by two parameters w and t_R , opening the door to a reliable estimation from data. Although the integration shape h could also be fit from data in theory, it seems more fruitful to assume a simple shape from the start (a classic square window kernel in our applications). Given that our goal is to estimate the coarse scales of percept formation, our method will likely be robust to various simple choices for h . As a simple example, we tested our method, assuming a square window kernel, on data produced by a readout using an exponential kernel, and still recovered the correct parameters w , t_R and K .

4.3.2 Non-deterministic extraction time

Our model, as presented above, makes another important assumption: that perceptual readout occurs at the same time t_R on every stimulus presentation. This assumption is likely to be valid in perceptual tasks that allow a fast reaction from the animal (‘reaction time’ tasks), in which case t_R will generally be as small as it can get (see, e.g., Stanford et al., 2010). However, when sensory stimulation lasts longer (say, over 500 msec) it opens the door to variations in t_R from trial to trial, or even to several reactualizations of percept f^* during the same trial. For example, imagine that the stimulus is a particular RGB color on a monitor, and you are asked to judge whether it contains more green (G) or blue (B). From intuition, we can tell that our performance in such a task will not sensibly increase whether we watch the color for one second or one minute. In our model’s formalism (eq. 6), this reveals built-in limitations on the effective integration window w that we can use in the task (remember that the readout’s performance is proportional to w). But then, if our percept arises from a limited integration window w and we indeed watch the color for a full minute, when is our percept built?

In appendix A, we derive a more general version of the characteristic equations (eq. 7-9) assuming that t_R in eq. 6 is itself a random variable, drawn on each trial following some probability distribution $g(t)$. Because sensory neurons have rather stationary activities in time, this additional assumption does not strongly affect the readout’s sensitivity. On the other hand, it affects strongly PCV curves. Essentially, the resulting PCV curve resembles a convolution of the deterministic PCV curve by $g(t)$ (Fig. 7b, section A.2.2). If $g(t)$ is substantially distributed in time, the PCV curves will become broader,

and flatter. In practice, this means that if a behavioral task is built such that t_R can display strong variations from trial to trial, the statistical method introduced above will produce biased estimates. In theory, this issue could be resolved by adding an additional parameter in the analysis to describe $g(t)$ (see section A.2.2), but the validation remains to be done.

4.3.3 Top-down influence of choice

Finally, our ‘standard’ model assumes that percept formation is exclusively feed-forward. The activities $s_i(t)$ of the sensory neurons are integrated to give rise to percept \hat{f} and the animal’s resulting choice c , and this forming decision does not affect sensory neurons in return. Recent evidence suggests that reality is more complex. By looking at the temporal evolution of CP signals in V2 neurons during a depth discrimination task, Nienborg and Cumming (2009) evidenced dynamics which are best explained by a top-down signal, biasing the activity of the neurons on each trial *after the choice is formed*. In our notations, the population spikes $s_i(t)$ would thus display a choice-dependent signal which kicks in on every trial after time t_R , resulting in PCV signals that deviate from their prediction in the absence of feedback (Fig. 7c).

What descriptive power does our model retain, if such top-down effects are strong? The answer depends on the nature of the putative feedback. If the feedback depends linearly on percept \hat{f} (and thus, on the spike trains), its effects are fully encompassed in our model. Indeed, this feedback signal will then be totally captured by the neurons’ linear covariance structure $\gamma_{ij}(t, s)$, so that our predictions will naturally take it into account. This is also the case if the oddity noted by Nienborg and Cumming (2009) is due to global shifts of neural excitability from trial to trial. On the other hand, if the feedback depends directly on the choice c —which displays a nonlinear, ‘all-or-none’ dependency on \hat{f} —then it will not be captured by our model, and lead to possible biases. Even so, the effects of the feedback could be largely alleviated through a small trick: compare true and predicted PCV signals only up to (candidate) time t_R (see eq. 22).

4.4 Extrapolation to larger neural ensembles

Can we understand in more depth the statistical principles at work underneath our method of estimation? What factors govern the evolution of sensitivity $Z(\mathcal{K})$ (Fig. 4a, eq. 14) and mean PCV signal $W(t|\mathcal{K})$ (Fig. 4b, eq. 16), as a function of the number of neurons K used for readout? This question is not only of theoretical, but also of practical interest. Indeed, it may happen in real applications that the number of simultaneously recorded neurons K_{\max} is too small to observe the crossing of predicted and true SNR curves (Fig. 4a)⁵. In such a case, predictions will be biased because no recorded ensemble \mathcal{K} can readily account for animal sensitivity.

What predictive power do ensembles up to size K_{\max} contain about larger ensembles? For example, can we extrapolate the shape of the mean SNR curve (Fig. 4a) to $K > K_{\max}$? In appendix B we address this question theoretically, by studying the value of SNR and PCV signals as a function of ensemble size K , and of the general structure of activity in the population. Our study relies on the singular value decomposition (SVD) of neural activity in the population. The SVD reveals a set of $m = 1 \dots M$ independent *modes* of population activity, each mode being associated to a power λ_m^2 and a sensitivity η_m^2 . Essentially, the sensitivity embedded in a neural ensemble \mathcal{K} of size K increases as the sum of sensitivities for the K first modes in the population—which are the modes with the largest powers λ_m . Conversely, the overall power of PCV signal $W(t)$ decreases as the average value of λ_m^2 in these K first modes, weighted by their respective sensitivities.

Because there is no general relationship between the power λ_m of a mode and its sensitivity to stimulus η_m , there is no trivial way of extrapolating SNR and PCV predictions to ensemble sizes K that were

⁵Actually, this is always bound to happen for small tested parameter w , following the K – w tradeoff.

not monitored simultaneously. Any such extrapolation can only be done through specific assumptions about the link between λ_m and η_m —which essentially amounts to characterizing the relative embedding of signal and noise in the full population (Wohrer et al., 2012). For example, it is classically assumed that the noise covariance matrix is “smooth” with respect to the signal covariance matrix, so that the former can be predicted on the basis of the latter (Wohrer et al., 2010; Haefner et al., 2013). Thus, while extrapolation of the statistical method above to larger populations is not trivial, it can be performed under specific assumptions about the embedding of signal and noise in the population considered.

5 Conclusion

We have shown how classic data recorded during perceptual decision-making experiments can be interpreted as samples from the joint covariance structure of neural activities and animal decision. Assuming a standard linear model of percept formation from neural activities, we derived a set of characteristic equations which relate neural and perceptual data, and thus define implicitly the parameters of perceptual integration by the animal on the basis of its sensory neurons. The neural data consist of neural PSTHs (first moment of neural activities) and JPSTHs (second moment of neural activities). The perceptual data consist of the animal’s sensitivity, and of each neuron’s covariance with the animal’s choice—a quantity often assessed through choice probabilities, and for which we proposed a simpler linear equivalent coined *percept covariance* (PCV).

We then proposed a method to utilize these characteristic equations in a case of practical interest, when experimenters only have access to finite statistical samples of neural data across the full population. Our goal was to successfully recover the instants in time and the typical number of neurons being used for percept formation—a difficult problem which cannot be solved on the sole basis of sensitivity information, due to the “ K – w tradeoff”. Our method relies on statistical averages of predicted sensitivity and PCV signals arising from random, candidate neural ensembles used as the source of percept formation ; and seeks to match these predictions with the true, recorded perceptual data. We tested this method on an artificial neural network producing a form of stimulus encoding, and showed that it successfully recovers the scales of perceptual integration, on the basis of sample recordings of realistic size.

Our method opens the way to novel experimental assessments of percept formation in sensory decision-making tasks. Indeed, the two main quantities used in our statistical analysis—sensitivity Z (eq. 14) and mean PCV curve $W(t)$ (eq. 16)—rely on classic experimental measures. The main limitation of our approach is the size of candidate readout ensembles which can be considered, as they should necessarily have been recorded simultaneously. However, the number of simultaneously recorded neurons is constantly pushing upwards with modern experimental techniques, so we may expect that this limitation, if it exists, will soon be overcome. Furthermore, through a theoretical analysis based on the singular value decomposition (SVD) of neural activities, we showed the possibility of extrapolation to larger ensemble sizes than those simultaneously recorded, although such extrapolations can only be done under specific assumptions, and on a case-by-case basis. For all these reasons, our method can readily be tested on real data, and hopefully provide new insights into the nature of percept formation from populations of sensory neurons.

Appendices

A Characteristic equations for the readout

A.1 Derivation

We here derive the characteristic equations for the linear readout introduced in the main text, and further comment some of its properties. We consider a more general version of eq. 6, where the extraction time t_R is allowed to vary from trial to trial. We thus assume that t_R is itself a random variable, drawn on each trial according to some density function $g(t)$, independently of neural activities $\mathbf{s}(t)$. The full readout model then writes:

$$\hat{f}(t_R) = \sum_i \int_{u>0} a^i s_i(t_R - u) h_w(u) du, \quad (\text{A.1})$$

$$t_R \sim g(t). \quad (\text{A.2})$$

This model naturally encompasses the simpler version presented in the main text, with a deterministic time t_R : it corresponds to taking $g(t)$ as a Dirac function located on that deterministic value.

The characteristic equations for this model rely on a straightforward computation of the second order statistics of \hat{f} , starting from eq. A.1. To deal with random time t_R , we note that for any random process $X(t)$ independent of t_R , $E(X(t_R)) = \int_{t=-\infty}^{+\infty} g(t) E(X(t)) dt$. This expression is valid only if t_R is independent from the random variables contributing to X (in our case, the spike trains).

Then, the expected value of \hat{f} given a stimulus f writes:

$$\begin{aligned} E(\hat{f}(t_R)|f) &= \int_t g(t) \sum_i \int_{u>0} a^i E(s_i(t - u)|f) h_w(u) du dt \\ &= \sum_i a^i \int_t (g \star h_w)(t) \lambda_i(t; f) dt, \end{aligned} \quad (\text{A.3})$$

where $g \star h_w(t) = \int_u g(u) h_w(u - t) du$ is the temporal correlation between g and h_w , and $\lambda_i(t; f)$ is the PSTH for neuron i in stimulus condition f , defined as in the main text (eq. 2).

Similarly, the expected value of \hat{f}^2 given a stimulus f writes:

$$\begin{aligned} E(\hat{f}(t_R)^2|f) &= \int_t g(t) E\left(\left(\sum_i \int_{u>0} a^i s_i(t - u) h_w(u) du\right)^2 \middle| f\right) dt \\ &= \int_t g(t) \sum_{ij} \iint_{(u,v)>0} a^i a^j E(s_i(t - u) s_j(t - v) | f) h_w(u) h_w(v) du dv dt \\ &= \sum_{ij} a^i a^j \iint_{(t,s)} G_w(t, s) \eta_{ij}(t, s; f) dt ds, \end{aligned} \quad (\text{A.4})$$

where we have defined $G_w(t, s) := \int_u g(u) h_w(u - t) h_w(u - s) du$, and $\eta_{ij}(t, s; f) := E(s_i(t) s_j(s) | f)$. $\eta_{ij}(t, s; f)$ is very related to the covariance structure in the population. It corresponds to the “plain” JPSTH for the neurons in stimulus condition f , before correcting by the so-called “product predictor” (Aertsen et al., 1989).

Finally, the expected value for the product of \hat{f} and the activity of any neuron $s_i(t)$ writes:

$$\begin{aligned} E(\hat{f}(t_R)s_i(t)|f) &= \int_s g(s) \sum_j \int_{u>0} a^j E(s_i(t)s_j(s-u)|f) h_w(u) du ds \\ &= \sum_j a^j \int_s (g \star h_w)(s) \eta_{ij}(t, s; f) ds, \end{aligned} \quad (\text{A.5})$$

using the same notations as above.

The three expressions eq. A.3-A.5 roughly correspond to the three characteristic equations for the readout. To obtain them, we consider the variational versions of the previous expressions. First, we obtain the characteristic equation for tuning by differentiating eq. A.3 with respect to stimulus. Second, equations A.4 and A.5 are expressed in ‘product’ form $E(XY)$, whereas the corresponding characteristic equations are expressed in ‘covariance’ form $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$. Once this is done, and after some rearrangement of the terms, we obtain the characteristic equations for tuning (eq. A.6), sensitivity (eq. A.7) and percept covariance (eq. A.8):

$$\partial_f E(\hat{f}|f) = \sum_i a^i \int_t (g \star h_w)(t) \beta_i(t) dt, \quad (\text{A.6})$$

$$\langle \text{Var}(\hat{f}|f) \rangle_f = \sum_{ij} a^i a^j \left(\iint_{(t,s)} G_w(t, s) \gamma_{ij}(t, s) dt ds + V_{ij}^{temp} \right), \quad (\text{A.7})$$

$$\langle \text{Cov}(\hat{f}, s_i(t)|f) \rangle_f = \sum_j a^j \int_s (g \star h_w)(s) \gamma_{ij}(t, s) ds. \quad (\text{A.8})$$

In eq. A.6, $\beta_i(t)$ is the temporal tuning curve for neuron i , defined as in eq. 3. If the readout is unbiased, the left-hand side is equal to 1, as in the main text. In eq. A.7 and A.8, $\gamma_{ij}(t, s)$ is the covariance structure (JPSTH) between neurons i and j , defined as in eq. 4.

Finally in eq. A.7, matrix V_{ij}^{temp} is an additional source of variance that appears only when $g(t)$ has an extended temporal support, i.e., when t_R is non-deterministic. It then writes:

$$V_{ij}^{temp} = \iint_{(t,s)} G_w(t, s) \left\langle \left(\lambda_i(t; f) - \bar{\lambda}_i(f) \right) \left(\lambda_j(s; f) - \bar{\lambda}_j(f) \right) \right\rangle_f dt ds,$$

where $\bar{\lambda}_i(f) := \int_t (g \star h_w)(t) \lambda_i(t; f) dt$ is the temporal average already used above (eq. A.3). Thus, V_{ij}^{temp} measures a form of temporal covariance in the PSTHs for the neurons.

When t_R is deterministic, as in the main text, we have $g(t) = \delta(t - t_R)$, a Dirac function. Then, the temporal integration kernels used in eq. A.6-A.8 boil down to $(g \star h_w)(t) = h_w(t_R - t)$ and $G_w(t, s) = h_w(t_R - t) h_w(t_R - s)$. One checks easily that in these conditions, the additional temporal variance term V_{ij}^{temp} vanishes, and we recover the characteristic equations from the main text.

A.2 Additional interpretations

A.2.1 Sensitivity as a function of w

In the form of eq. A.7, it is not clear how the value of w influences the variance of \hat{f} , and thus the readout’s sensitivity. To get a better intuition, let us first neglect the temporal variance term V_{ij}^{temp} . One checks easily that kernel $G_w(t, s)$, introduced above, verifies the following property: $\int_t G_w(t, t+\tau) dt =$

$(h_w \star h_w)(\tau)$, the autocorrelation of kernel h_w . As a result, we can rewrite A.7 in the form :

$$\begin{aligned} \langle \text{Var}(\hat{f}|f) \rangle_f &= \sum_{ij} a^i a^j \int_{\tau} (h_w \star h_w)(\tau) \left(\int_t \frac{G_w(t, t+\tau)}{h_w \star h_w(\tau)} \gamma_{ij}(t, t+\tau) dt \right) d\tau \\ &= \sum_{ij} a^i a^j \int_{\tau} (h_w \star h_w)(\tau) \overline{\gamma_{ij}}(\tau) d\tau. \end{aligned} \quad (\text{A.9})$$

In the first line, the function of t defined by the fraction is positive and has an integral of 1, so it operates as a temporal averaging on $\gamma_{ij}(t, t+\tau)$. The resulting average over t , noted $\overline{\gamma_{ij}}(\tau)$ in the second line, is thus a form of *cross-correlogram* between neurons i and j , measuring the average covariance between the spikes from i and j separated by a time lag τ .

Because h_w is a low-pass kernel with scale w , its autocorrelation function typically has support on $[-w, w]$, and verifies⁶: $(h_w \star h_w)(0) = w^{-1}$. On the other hand, $\overline{\gamma_{ij}}(\tau)$ typically has support on some interval $[-\tau_\gamma, \tau_\gamma]$, where τ_γ is the typical time scale of noise correlations in the population. As a result, as soon as w gets bigger than τ_γ , the integral in (A.9) starts behaving like w^{-1} , and the SNR of \hat{f} scales as w . A similar analysis can be performed on the additional term V_{ij}^{temp} (eq. A.7).

A.2.2 Non-deterministic t_R

What are the main departures from the main text when function $g(t)$ has an extended temporal support? From eq. A.6-A.8, it is clear that the general form of the characteristic equations still holds:

$$\begin{aligned} 1 &= \mathbf{b}^\top \mathbf{a}, \\ Z^{-1} &= \mathbf{a}^\top \mathbf{C} \mathbf{a}, \\ \boldsymbol{\pi}(t) &= \boldsymbol{\Gamma}(t) \mathbf{a}, \end{aligned}$$

but with more general definitions of $\mathbf{b}(w, g)$, $\mathbf{C}(w, g)$ and $\boldsymbol{\Gamma}(t|w, g)$. First, an additional covariance matrix \mathbf{V}^{temp} may contribute to \mathbf{C} , if neural activities are not stationary in time. Indeed, if t_R varies from trial to trial, any variation of firing rates in time creates an additional source of variability in \hat{f} .

Second, through eq. A.8, $g(s)$ acts a weighting factor over the PCV curves that would be obtained for each t_R : $\boldsymbol{\Gamma}(t|g) = \int_s g(s) \boldsymbol{\Gamma}(t|t_R = s) ds$. This leads to the spreading of PCV curves sketched in Fig. 7c.

These two features lead to lose one specific property of the deterministic case. When the “natural” temporal averaging of PCV signals was considered, that is $\bar{\boldsymbol{\pi}} = \int_t h_w(t_R - t) \boldsymbol{\pi}(t) dt$, the integrated PCV equation yielded $\bar{\boldsymbol{\pi}} = \mathbf{C} \mathbf{a}$, because $\bar{\boldsymbol{\Gamma}} = \mathbf{C}$. In the general case, the “natural” temporal averaging is $\bar{\boldsymbol{\pi}} = \int_t (g \star h_w)(t) \boldsymbol{\pi}(t) dt$, and one checks easily that $\bar{\boldsymbol{\Gamma}} \neq \mathbf{C}$. Thus, with general $g(t)$, the sensitivity (eq. A.7) and PCV (eq. A.8) equations become more dissociated.

In these conditions, it is unclear whether the statistical approach introduced in the main text could be extended, to also recover a non-deterministic extraction function $g(t)$. The main concern is that the temporal evolution of PCV signals is only determined by the aggregate function $(g \star h_w)(t)$ (eq. A.8), which cannot be used to disentangle $g(t)$ and w separately. However, general considerations suggest that the method could still work in that case. Indeed, the respective effects of $g(t)$ and w on the covariance structures used in eq. A.7-A.8 can roughly be thought of as a scaling:

$$\bar{\boldsymbol{\Gamma}} \simeq \epsilon(w, g) \mathbf{C}, \quad (\text{A.10})$$

because the overall “shape” of covariance between neurons (as opposed to its “strength”) does not depend much on the precise temporal integration used to compute their activity. Actually, under the specific

⁶ Assuming proper scaling for shape function h : $\int_u h(u) du = \int_u h(u)^2 du = 1$.

assumption that $\gamma_{ij}(t, s) = \overline{\gamma_{ij}}Q(|t - s|)$ (stationary activities with uniform temporal correlations), relationship (eq. A.10) can be shown to be exact, with

$$\epsilon(w, g) = \frac{\int_{\xi} \tilde{Q}(\xi) \|\tilde{h}_w(\xi)\|^2 \|\tilde{g}(\xi)\|^2 d\xi}{\int_{\xi} \tilde{Q}(\xi) \|\tilde{h}_w(\xi)\|^2 d\xi}$$

expressed in terms of Fourier transforms. As a result, the mean PCV curve $W(t)$ (eq. 15, 16) is predicted to scale as $\epsilon(w, g)$. So, while matching the temporal support of $W(t)$ and $W^*(t)$ constrains the value of $(g \star h_w)(t)$, matching their overall power constrains $\epsilon(w, g)$, and we can hope to disentangle the values of $g(t)$ and w separately. In practice though, this would require the fitting of at least one additional temporal parameter ; typically, the standard deviation of t_R from trial to trial.

B Singular value analysis—summary

We summarize here the main results of a theoretical analysis to understand the evolution of SNR and PCV signals achieved by readout ensembles of growing size K . Detailed mathematical derivations are available in Supplementary Section S2. For simplicity we focus only on time-integrated neural activities $\bar{s}_i := \int_u h_w(u) s_i(t_R - u) du$, assuming a fixed choice of (w, t_R) . We consider random readout ensembles \mathcal{K} in the population, and two resulting indicators. First, we consider the sensitivity $Y(\mathcal{K})$, linked to SNR Z by relationship $Y = Z(1 + Z)^{-1}$. This is the natural description of sensitivity in the framework below. It is obtained like Z in the main text (eq. 14) but replacing the noise covariance matrix \mathbf{C} by the total covariance matrix $\mathbf{A} = \mathbf{C} + \langle f^2 \rangle \mathbf{b}\mathbf{b}^\top$. Second, we consider the mean PCV in the population $\bar{W}(\mathcal{K})$, obtained as the “natural” temporal integration of signal $W(t|\mathcal{K})$ from the main text (eq. 16): $\bar{W} := \int_u h_w(u) W(t_R - u) du$. Since $W(t)$ is mostly positive, \bar{W} roughly corresponds to the overall power in $W(t)$.

SVD reformulation of neural activity. The analysis relies on the singular value decomposition (SVD) of population activity into $m = 1 \dots M$ orthogonal *modes*:

$$\bar{s}_i^{f\omega} = \sum_{m=1}^M \lambda_m u_i^m v_m^{f\omega},$$

where the lower index $i = 1 \dots N_{\text{tot}}$ indicates neurons in the population, and the upper index indicates all possible stimuli f and random realizations ω of network activity. Each mode m is defined by its power $\lambda_m > 0$, its distribution vector (over neurons) \mathbf{u}^m , and its appearance variable v_m which takes a different random value on every trial. By construction, the various modes are orthogonal across neurons $((\mathbf{u}^m)^\top \mathbf{u}^n = \delta^{mn})$, and linearly independent across trials $(\text{Cov}_{f\omega}(v_m, v_n) = \delta_{mn})$, so they typically correspond to distinct “patterns of activity” in the population. The power λ_m describes the overall impact of mode m on population activity. We assume $\lambda_1 \geq \dots \geq \lambda_M$, so we progressively include modes with lower power—either because they involve only a small fraction of neurons, either because they appear only on rare trials. The number of modes M is the intrinsic dimensionality of the neural population’s activity. In real populations we expect $M \ll N_{\text{tot}}$, because neural activities are largely correlated.

The SVD is best viewed as a change of variables reexpressing neural activities $\{\bar{s}_i\}_{i=1 \dots N_{\text{tot}}}$ in terms of mode appearance variables $\{v_m\}_{m=1 \dots M}$. Just like individual neurons, each mode m can be associated to a *sensitivity* to stimulus η_m , which describes the proportion of the mode’s power λ_m due to variations of the signal (f), as opposed to variations of the noise (ω). Since modes are linearly independent, the full population’s sensitivity corresponds to the sum of individual mode sensitivities: $Y(\infty) = \sum_m \eta_m^2$.

Sensitivity and PCV from finite neural ensembles. We now want to estimate the amount of stimulus sensitivity $Y(\mathcal{K})$ that can be extracted, not from the full population, but from neural subensembles of size \mathcal{K} . The SVD provides a natural reinterpretation of this problem in terms of activity modes: each ensemble \mathcal{K} “reveals” only a fraction of the underlying modes. The pivotal object to perform this reinterpretation is our so-called *data matrix*:

$$\mathbf{D}_{\mathcal{K}} := \left\{ d_i^m = \lambda_m u_i^m \right\}_{i \in \mathcal{K}}^{m=1 \dots M},$$

an $M \times K$ matrix describing the activity of neural ensemble \mathcal{K} in the space of modes. In the original problem formulation, the $K \times K$ matrix $\mathbf{D}^\top \mathbf{D}$ describes the covariance of neural activity in ensemble \mathcal{K} , and we want to estimate the resulting sensitivity. In the dual formulation, the $M \times M$ matrix $\mathbf{D} \mathbf{D}^\top$ describes a covariance structure between modes, but estimated only from the sample neurons in \mathcal{K} . The problem now lives in a space of fixed dimensionality M , and can be related to classical problems of estimating covariance structures from a finite number of samples—in our case, the neurons.

Applying this dual approach, we find that $Y(\mathcal{K})$ and $\bar{W}(\mathcal{K})$ depend on readout ensemble \mathcal{K} only through an $M \times M$ matrix $\Delta_{\mathcal{K}}$, the (rank K) orthogonal projector on the span of vectors $\{\mathbf{d}_i\}_{i \in \mathcal{K}}$ in mode space:

$$Y(\mathcal{K}) = \boldsymbol{\eta}^\top \Delta_{\mathcal{K}} \boldsymbol{\eta},$$

$$\bar{W}(\mathcal{K}) = -B^2 + (N_{\text{tot}} Y(\mathcal{K}))^{-1} \boldsymbol{\eta}^\top \mathbf{\Lambda}^2 \Delta_{\mathcal{K}} \boldsymbol{\eta},$$

where $B^2 := \mathbb{E}_i(b_i^2)$ is the average square tuning in the population. Furthermore, the average projector $\Delta_{\mathcal{K}}$ across ensembles of size K , that is $\mathbb{E}_K \Delta$, is approximately diagonal in mode space. Noting $\{\epsilon_K^m\}$ for its diagonal, we thus obtain the approximations:

$$\mathbb{E}_K Y \simeq \sum_{m=1}^M \epsilon_K^m \eta_m^2, \quad (\text{B.1})$$

$$\mathbb{E}_K \bar{W} \simeq -B^2 + N_{\text{tot}}^{-1} \frac{\sum_{m=1}^M \epsilon_K^m \eta_m^2 \lambda_m^2}{\sum_{m=1}^M \epsilon_K^m \eta_m^2}, \quad (\text{B.2})$$

where $0 \leq \epsilon_K^m \leq 1$ is the average “proportion” of mode m revealed by K random neurons. As modes with larger power λ_m tend to be revealed first, a rough but useful image is to consider that $\epsilon_K^m \simeq \mathbb{1}_{m \leq K}$ —only the K first modes are revealed by ensembles of K neurons.

Thus, sensitivity $\mathbb{E}_K Y$ grows with K as mode sensitivities η_m are progressively revealed. Saturation occurs when all nonzero η_m are revealed, in which case $\mathbb{E}_K Y = Y(\infty)$. Conversely, the mean PCV $\mathbb{E}_K \bar{W}$ decreases with K . Indeed, the fraction in eq. B.2 can be viewed as an average power $\langle \lambda^2 \rangle_{m,K}$, where each mode m contributes with a weight $\epsilon_K^m \eta_m^2$. As ϵ_K^m progressively reveals modes with lower power λ_m , this average power is expected to decrease with K . Again, saturation occurs when all nonzero η_m are revealed, and then $\mathbb{E}_K \bar{W} = Z(\infty)^{-1} B^2$, the predicted value for choice signals in case of optimal readout from the full population (Haefner et al., 2013).

Extrapolation to large K . What do these results tell us about possible extrapolations to ensemble sizes K larger than the maximum number of neurons simultaneously recorded by the experimenter? Essentially, that such extrapolations always require further assumptions about the structure of activity in the population.

Indeed, one can imagine scenarios in which the most sensitive modes (those with highest η_m^2) are associated to relatively low powers λ_m^2 and thus, appear only at large K . This could be the case, for example, if a very local circuit of neurons carries a lot of information about the stimulus, independently from the rest of the population. Because it involves few neurons, the corresponding mode of activity

will have a low power λ_m^2 , and will require very large ensembles \mathcal{K} to be detected—simply because the corresponding neurons are not recorded in smaller ensembles. A similar discussion can be found in Haefner et al. (2013). Another example is the encoding network theoretically proposed by Boerlin and Denève (2011), where each neuron spikes only if its information is not already encoded in the activity of the remaining neurons. This results in the appearance of a few, global modes of activity⁷ which are specifically designed to have a very large SNR, meaning high η_m and low λ_m . In this case, any estimation of sensitivity from a subpopulation \mathcal{K} will consistently be smaller than the full population’s sensitivity.

To summarize, extrapolation can only be performed under additional assumptions about the overall link between η_m and λ_m —or equivalently, about the relationship between ‘signal’ and ‘noise’ contributions to population activity (see also discussions in Wöhrer et al., 2012; Haefner et al., 2013). The extent to which such assumptions are justified will depend on each specific context.

References

- Abbott, L. F., Dayan, P., 1999. The effect of correlated variability on the accuracy of a population code. *Neural computation* 11 (1), 91–101.
- Aertsen, A. M., Gerstein, G. L., Habib, M. K., Palm, G., 1989. Dynamics of neuronal firing correlation: modulation of “effective connectivity”. *Journal of neurophysiology* 61 (5), 900–17.
- Averbeck, B. B., Latham, P. E., Pouget, A., 2006. Neural correlations, population coding and computation. *Nature Reviews Neuroscience* 7 (5), 358–66.
- Boerlin, M., Denève, S., 2011. Spike-based population coding and working memory. *PLoS computational biology* 7 (2).
- Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., Movshon, A. J., 1996. A relationship between behavioral choice and the visual response of neurons in macaque MT. *Visual Neuroscience* 13, 87–100.
- Britten, K. H., Shadlen, M. N., Newsome, W. T., Movshon, J. A., 1992. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience* 12 (12), 4745–4765.
- Cohen, M. R., Newsome, W. T., 2009. Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *Journal of Neuroscience* 29 (20), 6635–48.
- Daley, D., Vere-Jones, D., 2007. An introduction to the theory of point processes. Vol. 1. Springer Verlag, New York, USA.
- de Lafuente, V., Romo, R., 2006. Neural correlate of subjective sensory experience gradually builds up across cortical areas. *Proceedings of the National Academy of Sciences of the United States of America* 103 (39), 14266–71.
- Gold, J. I., Shadlen, M. N., 2007. The neural basis of decision making. *Annual Review of Neuroscience* 30, 535–74.
- Goodman, D., Brette, R., 2008. Brian: a simulator for spiking neural networks in python. *Frontiers in neuroinformatics* 2.
- Green, D., Swets, J., 1966. Signal detection theory and psychophysics. Vol. 1974. Wiley, New York, USA.
- Haefner, R. M., Gerwinn, S., Macke, J. H., Bethge, M., 2013. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nature Neuroscience* 16 (2), 235–242.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Springer Verlag, New York, USA.

⁷Typically the sum of all neural activities, in the simplest instantiation of the model.

- Hernández, A., Zainos, A., Romo, R., 2000. Neuronal correlates of sensory discrimination in the somatosensory cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97 (11), 6191–6.
- Hoyle, D. C., 2011. Accuracy of pseudo-inverse covariance learning—a random matrix theory analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33 (7), 1470–1481.
- Luna, R., Hernández, A., Brody, C. D., Romo, R., 2005. Neural codes for perceptual discrimination in primary somatosensory cortex. *Nature Neuroscience* 8 (9), 1210–9.
- Mountcastle, V., Steinmetz, M. A., Romo, R., 1990. Frequency discrimination in the sense of flutter: psychophysical measurements correlated with postcentral events in behaving monkeys. *Journal of Neuroscience* 10 (9), 3032–3044.
- Nienborg, H., Cumming, B. G., 2009. Decision-related activity in sensory neurons reflects more than a neuron’s causal effect. *Nature* 459 (7243), 89–92.
- Price, N. S. C., Born, R. T., Oct. 2010. Timescales of sensory- and decision-related activity in the middle temporal and medial superior temporal areas. *Journal of Neuroscience* 30 (42), 14036–45.
- Raudys, S., Duin, R., 1998. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters* 19 (5), 385–392.
- Romo, R., Salinas, E., 2003. Flutter discrimination: neural codes, perception, memory and decision making. *Nature Reviews Neuroscience* 4 (3), 203–18.
- Shadlen, M. N., Britten, K. H., Newsome, W. T., Movshon, A. J., 1996. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience* 16 (4), 1486–1510.
- Shadlen, M. N., Newsome, W. T., 1998. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience* 18 (10), 3870–3896.
- Stanford, T. R., Shankar, S., Massoglia, D. P., Costello, M. G., Salinas, E., 2010. Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience* 13 (3), 379–385.
- Talbot, W., Darian-Smith, I., Kornhuber, H., Mountcastle, V., 1968. The sense of flutter-vibration: comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. *Journal of Neurophysiology* 31 (2).
- Uka, T., DeAngelis, G. C., 2003. Contribution of middle temporal area to coarse depth discrimination: comparison of neuronal and psychophysical sensitivity. *Journal of Neuroscience* 23 (8), 3515–30.
- Werner, G., Mountcastle, V., 1965. Neural activity in mechanoreceptive cutaneous afferents: Stimulus-response relations, weber functions, and information transmission. *Journal of Neurophysiology* 28 (2).
- Wohrer, A., Humphries, M. D., Machens, C., 2012. Population-wide distributions of neural activity during perceptual decision-making. *Progress in Neurobiology*.
- Wohrer, A., Romo, R., Machens, C., 2010. Linear readout from a neural population with partial correlation data. No. 1. pp. 2–10.

SUPPLEMENTARY INFORMATION

S1 Encoding network

We detail here the architecture of the artificial encoding network used to test our method (summarized in section 3.1 from the main text). This ad-hoc network was designed to display some classic features of sensory cortical neurons involved in perceptual decision-making tasks (e.g, V2, MT, S1, S2...). To reproduce the diversity of response naturally observed at the population level (Wohrer et al., 2012), neurons in our network have broadly distributed firing rates, and some diversity in their temporal response profiles. We also wished to reproduce the continuum of tuning to stimulus observed in real populations, where some neurons have positive tuning to stimulus (rate increase when f increases), and other neurons have negative tuning. Finally, we wished to reproduce realistic strengths of noise correlations between neurons in the population (Figure 3b from the main text), and insure that the tunings of each pair of neurons (their “signal” correlation) be only slightly predictive of their noise correlation—another feature often observed in real sensory populations (Wohrer et al., 2012).

The network consists of two distinct layers of spiking neurons, of which only the second layer (encoding layer) is “visible” to the experimenter. The first layer (L1) consists of $100 = 2 \times 50$ independent Poisson neurons, whose firing intensity f constitutes the stimulus encoded by the second layer. On each trial, f takes one of three possible values 25, 30 and 35 Hz. All neurons are equivalent, but segregated in two distinct populations according to their projections on the second layer. The Poisson firing constitutes the only source of randomness in the network from trial to trial.

The second layer (L2) consists of 500 leaky integrate-and-fire (LIF) neurons, some of which receive input from L1, and who are all coupled through a sparse, balanced connectivity. The generic equation for these neurons writes

$$\tau \frac{dV_i^{(s)}}{dt} = \sum_{j \in \text{L1}} W_{ji}^{(1,s)} \delta(t - t_j) + \sum_{k \in \text{L2}} W_{ki}^{(2)} \delta(t - t_k - \Delta_{ki}) + I^{(s)} - (V_i^{(s)}(t) - V^{rest}).$$

The neuron emits a spike at each time t_i when $V_i^{(s)}$ reaches threshold V^{thr} , after what the neuron’s potential is reinitialized at resting value V^{rest} . All neurons share the same membrane time constant $\tau = 20$ msec, threshold $V^{thr} = -50$ mV, and resting potential $V^{rest} = -60$ mV. Upper index s denotes one of three possible subtypes of neurons in L2: Positively-biased neurons ($s = p$, 100 neurons), negatively-biased neurons ($s = n$, 100 neurons) and unbiased neurons ($s = u$, 300 neurons).

Positively-biased neurons receive sparse excitatory connections from 50 neurons in L1 ($W_{ji}^{(1,p)} \geq 0$), whereas negatively-biased neurons receive sparse inhibitory connections from the 50 other neurons in L1 ($W_{ji}^{(1,n)} \leq 0$). Unbiased neurons receive no direct input from L1 ($W_{ji}^{(1,u)} = 0$). As these asymmetries create biases in the total synaptic inputs to each type of cell, the intrinsic currents $I^{(p)}$, $I^{(n)}$ and $I^{(u)}$ also vary depending on neuron subtype, to insure homogeneous firing properties inside the three populations (see Table 1). Finally, all L2 neurons are connected through a single matrix $\mathbf{W}^{(2)}$ of recurrent connections—independently of their subtype. All connection matrices $\mathbf{W}^{(1,s)}$ and $\mathbf{W}^{(2)}$ are sparse with (Erdős-Renyi) connection probability $p = 0.2$. Non-zero connection strengths are picked uniformly in an interval $[w_{\min}, w_{\max}]$, which depends on the connection considered: see Table 1. Note that L2 recurrent connections can be both excitatory and inhibitory, a departure from biology which allows for an easier implementation.

Finally, the recurrent connections in L2 are associated to synaptic delays: for each pair (i, k) of connected L2 neurons, the random delay Δ_{ki} is drawn uniformly between 0 and 5 msec. This substantially increases the diversity of neural responses in the population, particularly at the level of JPSTHs (Figure 3e from the main text)—this is interesting because our method is specifically designed to analyse generic,

Subtype	$I^{(s)}$	$w_{\min}^{(1,s)}$	$w_{\max}^{(1,s)}$	$w_{\min}^{(2)}$	$w_{\max}^{(2)}$
Pos. biased (p)	0	0	2	-2	2
Neg. biased (n)	14	-3	0	-2	2
Unbiased (u)	5	0	0	-2	2

Table 1. Connectivity parameters in the three subtypes of L2 neurons. All values are expressed in millivolts.

heterogeneous population activities.

We implemented and simulated the network using Brian, a spiking neural network simulator in Python (Goodman and Brette, 2008). Our simulation consisted of many successive epochs of 500 msec with all possible successions of the three stimulus values f (as in Figure 1a from the main text). Since the input Poisson neurons were always firing close to 30 Hz, there was no strong transient at stimulus onset as is often observed in real sensory neurons. In our case, the change of activity between two successive stimuli was always only differential, and rather weak (see Figure 3c from the main text).

S2 Singular value analysis

We detail here our mathematical analysis to understand the evolution of SNR and PCV estimates in growing populations of size K , as a function of the underlying structure of the full population. These results expand the condensed presentation proposed in appendix B of the main text.

1 Notations

1.1 Activity across neurons, stimuli and trials

For simplicity, we consider a timeless version of neural activities, although the whole analysis could be extended to include time as well. In our readout framework, this means that we fix some candidate temporal integration parameters (w, t_R) , and consider the resulting neural activities S_i , constructed from the temporal integration of each neuron i 's spikes¹.

Since our main results have been presented in the case of linear tuning to stimuli, we stick to this hypothesis. This implies that all signal/noise properties can be understood by considering only two stimuli (as the difference in response between these two stimuli totally defines the linear tuning of each neuron). We thus note $f = \{0, 1\}$ the two possible stimulus values which can be input to the network.

Finally, we may want to consider the possibility of imprecise neural measurements, due to recording from only a finite number of trials (although it is not the main concern of this note). We thus denote $\omega \in \Omega$ the set of all possible different realizations of network activity. In theory, Ω is an infinite set of possible events. However, we will formally assume it to be finite, with (huge) cardinality Ω —so on a given trial, each possible network realization ω has a probability $1/\Omega$ of coming out.

We thus summarize all possible network realizations through the array $S_i^{f\omega}$, where $i = 1 \dots N$ denotes all neurons in the population², $f = 0, 1$ denotes stimulus value, and $\omega = 1 \dots \Omega$ denotes all possible realizations. The notation $f\omega$, somewhat abusive, applies the same indexing ω for possible realizations in both stimulus conditions $f = 0$ and $f = 1$ —which can only be done if both stimulus conditions allow the same number Ω of possible network realizations. However, given the formal nature of ensemble Ω , this notation abuse appears harmless.

As we start doing statistics across neurons and trials, we will need to compute expectancies (i.e., means) and covariance structures across various dimensions. In all cases, we apply the generic notation $E_\alpha^A(X_{\alpha,\beta,\dots})$ to denote the empirical mean of quantity $X_{\alpha,\beta,\dots}$ when α is varied over ensemble A (β, \dots being any other parameters that are held fixed). When ensemble A is unambiguous, meaning that it includes all possible values for α , we will omit it. Finally, second order variances and covariance structures will generically be computed as $\text{Cov}_\alpha^A(X_\alpha, Y_\alpha) = E_\alpha^A(X_\alpha Y_\alpha) - E_\alpha^A(X_\alpha)E_\alpha^A(Y_\alpha)$.

As a first application of these notations, remember that the whole sensitivity analysis derived in the main text deals only with variations: the “signal” measures variations of activity with a change in stimulus f , while the “noise” measures variations of activity across trials ω . Thus, the overall mean level of activity for each neuron i , that is $E_{f\omega}(S_i^{f\omega})$, plays no role in the analysis: it always disappears from the computations of tuning and noise covariance structure. To clarify further notations, we can thus offset all neural signals and assume that $E_{f\omega}(S_i^{f\omega}) = 0$, for every neuron i in the population.

1.2 Modes of activity in the neural population

The key argument of this note relies on interpreting $S_i^{f\omega}$ as a very large $N \times (2\Omega)$ matrix, and considering its singular value decomposition (SVD). The (compact) SVD is a standard decomposition which can be applied to any rectangular matrix \mathbf{S} . It writes $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda}$ is an $M \times M$ diagonal matrix with

¹ S_i is noted \bar{s}_i in the main text.

² N is noted N_{tot} in the main text.

strictly positive entries λ_m (the singular values), \mathbf{U} is an $N \times M$ matrix of orthogonal columns (meaning $\mathbf{U}^\top \mathbf{U} = \mathbf{Id}_M$), and \mathbf{V} is an $\Omega \times M$ matrix of orthogonal columns (meaning $\mathbf{V}^\top \mathbf{V} = \mathbf{Id}_M$).

With our current definition of neural activity S , the SVD decomposition writes

$$S_i^{f\omega} = \sum_{m=1}^M \lambda_m u_i^m v_m^{f\omega}, \quad (1)$$

where the orthogonality of \mathbf{U} writes:

$$\forall (m, n), \sum_{i=1}^N u_i^m u_i^n = \delta^{mn}, \quad (2)$$

and the orthogonality of \mathbf{V} similarly writes $\sum_{f\omega} (v_m^{f\omega} v_n^{f\omega}) = \delta_{mn}$. In the case of \mathbf{V} , our above convention that $E_{f\omega}(S_i^{f\omega}) = 0$ for all neurons i actually imposes that $E_{f\omega}(v_m^{f\omega}) = 0$ for all modes m . We thus reinterpret the orthogonality of \mathbf{V} as a linear independence between the different random variables v_m :

$$\forall m, E_{f\omega}(v_m^{f\omega}) = 0, \quad (3)$$

$$\forall (m, n), \text{Cov}_{f\omega}(v_m^{f\omega}, v_n^{f\omega}) = \delta_{mn}. \quad (4)$$

Note that we reinterpret the sum over trials (f, ω) as an expectancy (thus rescaling λ_m by ensemble size 2Ω). This allows to emphasize the statistical interpretation of the SVD decomposition in this case.

Each triplet $(\lambda_m, \mathbf{u}^m, v_m)$ defines one particular *mode* of activity in the population. We call λ_m the *power* of the mode, \mathbf{u}^m (viewed as an N -dimensional vector) its *distribution vector*, and v_m (viewed as a scalar random variable) its *appearance variable*. The appearance variable v_m —which takes a different value $v_m^{f\omega}$ on every repetition of the experiment—describes the probability of appearance of each mode m across stimuli and trials. Through eq. 4, each mode m verifies $E_{f\omega}((v_m^{f\omega})^2) = 1$, meaning that all modes have the same overall “expected appearance” across trials.

Similarly, eq. 2 implies that $\sum_i ((u_i^m)^2) = 1$, so \mathbf{u}^m describes the normalized distribution of the mode across the neural population. Some modes m may correspond to a rather homogeneous distribution of $(u_i^m)^2$ across the population, meaning that the mode is very *distributed*, whereas other modes may have power concentrated only over a small subensemble of neurons. These are the modes corresponding to local patterns of activity which only impact a small fraction of the total neural population.

Finally, the power λ_m describes the overall impact of mode m on population activity. Indeed, although distribution vectors \mathbf{u}^m and appearance variables v_m display the same normalization across modes, this does not mean that all modes are equivalent. Instead, only those modes with the largest values λ_m will truly impact the population, in the form of measurable changes of activity across neurons and trials. Conversely, modes with small values λ_m will scarcely impact population activity, either because they involve only a small fraction of neurons, either because they are distributed but very weak.

The overall number of modes M is equal to the rank of matrix $S_i^{f\omega}$, so it is by construction smaller or equal to the population size N (which we assume to be smaller than the huge number Ω of possible realizations across trials). M defines the typical dimension of the manifold in which all neural activity occurs. In real neural populations, although N is itself a very large number, there are reasons to believe that M is sensibly smaller, due to correlated activity between neurons.

1.3 Statistics of activity

We now reinterpret classical measures of neural activity in the framework defined above. At this point, we need to carefully specify the nature of the ensembles truly available for measures: a finite subset \mathcal{K} of

neurons from the population, and a finite ensemble \mathcal{E} of trials (each element of \mathcal{E} providing one realization for stimulus $f = 0$ and one realization for $f = 1$).

For every neuron $i \in \mathcal{K}$, recorded over trials $\omega \in \mathcal{E}$, we compute the tuning to stimulus as

$$b_i^\mathcal{E} := \frac{1}{2} \left(\mathbb{E}_\omega^\mathcal{E}(S_i^{1\omega}) - \mathbb{E}_\omega^\mathcal{E}(S_i^{0\omega}) \right), \quad (5)$$

that is, the difference between the experimental mean firing rates in stimulus conditions $f = 1$ and $f = 0$.

³ Similarly, we compute the noise covariance term between any two neurons i and j as:

$$C_{ij}^\mathcal{E} := \frac{1}{2} \left(\text{Cov}_\omega^\mathcal{E}(S_i^{0\omega}, S_j^{0\omega}) + \text{Cov}_\omega^\mathcal{E}(S_i^{1\omega}, S_j^{1\omega}) \right), \quad (6)$$

that is, the stimulus-averaged noise covariance between i and j . Finally, we introduce the total covariance matrix $A_{ij}^\mathcal{E}$ summing up all sources of variance across the population:

$$\begin{aligned} A_{ij}^\mathcal{E} &:= \text{Cov}_{f\omega}^\mathcal{E}(S_i^{f\omega}, S_j^{f\omega}) \\ &= C_{ij}^\mathcal{E} + b_i^\mathcal{E} b_j^\mathcal{E}. \end{aligned} \quad (7)$$

The last line provides the classic decomposition of the total covariance matrix into noise covariance matrix $\mathbf{C}^\mathcal{E}$ and signal covariance matrix $(\mathbf{b}^\mathcal{E})(\mathbf{b}^\mathcal{E})^\top$ —which has rank 1 under our assumption of linear tuning to stimulus.

When ensemble \mathcal{E} is equal to the full space Ω of possible realizations, the above formulas define the “true” measures of covariance, as would be obtained given a sufficient amount of trials. In the sequel, we refer to these true, error-free values, by removing the mention to \mathcal{E} . That is: b_i , C_{ij} and A_{ij} .

The SVD decomposition (eq. 1) is best interpreted as a change of variables reexpressing neural activities $\{S_i\}_{i=1\dots N}$ in terms of mode appearance variables $\{v_m\}_{m=1\dots M}$. As a result, we can define the respective equivalents of tuning, noise covariance and total covariance in the space of activity modes. Indeed, although mode appearance variables v_m are never directly observed, they still have some statistics across trials. We thus define:

$$\begin{aligned} \eta_m^\mathcal{E} &:= \frac{1}{2} \left(\mathbb{E}_\omega^\mathcal{E}(v_m^{1\omega}) - \mathbb{E}_\omega^\mathcal{E}(v_m^{0\omega}) \right), \\ \Phi_{mn}^\mathcal{E} &:= \text{Cov}_{f\omega}^\mathcal{E}(v_m^{f\omega}, v_n^{f\omega}), \end{aligned}$$

which define tuning and total covariance in mode space (noise covariance being implicitly defined as $\Phi^\mathcal{E} - (\eta^\mathcal{E})(\eta^\mathcal{E})^\top$). Again, we will denote the true tuning and covariance by removing the mention to \mathcal{E} : true tuning $\boldsymbol{\eta}$ and true total covariance Φ . Importantly, the normalization of variables v_m in eq. 4 implies that $\Phi = \mathbf{Id}_M$.

Mode powers λ_m and distribution vectors \mathbf{u}^m then allow to relate the statistics at the levels of neurons and modes. Injecting the SVD formula (eq. 1) into equations 5 and 7 yields (in matricial form):

$$\mathbf{b}^\mathcal{E} = \mathbf{U} \boldsymbol{\Lambda} \boldsymbol{\eta}^\mathcal{E}, \quad (8)$$

$$\mathbf{A}^\mathcal{E} = \mathbf{U} \boldsymbol{\Lambda} \Phi^\mathcal{E} \boldsymbol{\Lambda} \mathbf{U}^\top. \quad (9)$$

In particular, when true noiseless measures are considered so that $\Phi = \mathbf{Id}_M$, we see that \mathbf{U} and $\boldsymbol{\Lambda}$ directly provide the standard (nonzero) eigenvalue decomposition of the total covariance matrix \mathbf{A} , as

$$\mathbf{A} = \mathbf{U} \boldsymbol{\Lambda}^2 \mathbf{U}^\top.$$

³Vector \mathbf{b} from this appendix corresponds to $\sigma_f \mathbf{b}$ from the main text, where $\sigma_f^2 = \langle f^2 \rangle_f - \langle f \rangle_f^2$ gives typical variations of input stimulus.

2 SNR and PCV predictions

We now wish to understand which factors determine the evolution of curve $Z(K)$, the average SNR embedded in neural subensembles \mathcal{K} of cardinal K . We can also study the evolution of percept covariance (PCV) signals, in the same framework.

In the main text, we compute SNR and PCV for ensemble \mathcal{K} through Fisher's linear discriminant (eq. 13-16). One sees easily that these definitions, involving tuning \mathbf{b} and noise covariance matrix \mathbf{C} , are equivalently expressed in terms of tuning \mathbf{b} and *total* covariance matrix \mathbf{A} :

$$\mathbf{a}_{\mathcal{K}} = (\mathbf{b}_{\mathcal{K}}^{\top} \mathbf{A}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}})^{-1} \mathbf{A}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}}, \quad (10)$$

$$Y(\mathcal{K}) = \mathbf{b}_{\mathcal{K}}^{\top} \mathbf{A}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}}. \quad (11)$$

We call Y the signal-to-total ratio (STR), which relates directly to SNR Z by the formula $Y = Z/(1+Z)$. Y always takes values between 0 ($Z = 0$) and 1 ($Z = \infty$), it thus avoids singularities which may occur in the direct Z formulation. If matrix $\mathbf{A}_{\mathcal{K}}$ is rank-deficient, we consider its (Moore-Penrose) pseudoinverse without loss of generality (see further down).

2.1 Total STR in the population

The SVD decomposition (eq. 1) reexpresses neural activity in the space of modes $m = 1 \dots M$. When the full neural population is considered, the full matrix \mathbf{A} and vector \mathbf{b} are involved in eq. 11. Using the SVD formulations (eq. 8-9) we thus find:

$$\begin{aligned} Y(\infty) &= \mathbf{b}^{\top} \mathbf{A}^{-1} \mathbf{b} \\ &= \boldsymbol{\eta}^{\top} \boldsymbol{\Lambda} \mathbf{U}^{\top} (\mathbf{U} \boldsymbol{\Lambda}^2 \mathbf{U}^{\top})^{-1} \mathbf{U} \boldsymbol{\Lambda} \boldsymbol{\eta} \\ &= \|\boldsymbol{\eta}\|^2 = \sum_{m=1}^M \eta_m^2. \end{aligned} \quad (12)$$

Thus, each mode contributes to total sensitivity by the strength of its intrinsic sensitivity η_m .

This computation can also be derived assuming a finite number of experimental trials \mathcal{E} . In this case however, we must introduce the *experimental sensitivity* $\zeta_m^{\mathcal{E}}$ of each mode, defined as

$$\boldsymbol{\zeta}^{\mathcal{E}} := (\boldsymbol{\Phi}^{\mathcal{E}})^{-\frac{1}{2}} \boldsymbol{\eta}^{\mathcal{E}}, \quad (13)$$

where $(\boldsymbol{\Phi}^{\mathcal{E}})^{-\frac{1}{2}}$ is the unique (Moore Penrose) pseudo-inverse of the symmetric, non-negative square root matrix of $\boldsymbol{\Phi}^{\mathcal{E}}$. Actually, any other choice of matrix square root could also be used, because by construction $\boldsymbol{\Phi}^{\mathcal{E}} \succeq (\boldsymbol{\eta}^{\mathcal{E}})(\boldsymbol{\eta}^{\mathcal{E}})^{\top}$, in the sense of symmetric positive matrices. This insures that $\boldsymbol{\eta}^{\mathcal{E}}$ is orthogonal to $\text{Ker}(\boldsymbol{\Phi}^{\mathcal{E}})$, and thus the unicity of $\boldsymbol{\zeta}^{\mathcal{E}}$ as defined in eq. 13.

The computation of $Y(\infty, \mathcal{E})$ then goes along the same lines as previously:

$$\begin{aligned} Y(\infty, \mathcal{E}) &= (\mathbf{b}^{\mathcal{E}})^{\top} (\mathbf{A}^{\mathcal{E}})^{-1} \mathbf{b}^{\mathcal{E}} \\ &= (\boldsymbol{\eta}^{\mathcal{E}})^{\top} \boldsymbol{\Lambda} \mathbf{U}^{\top} (\mathbf{U} \boldsymbol{\Phi}^{\mathcal{E}} \boldsymbol{\Lambda} \mathbf{U}^{\top})^{-1} \mathbf{U} \boldsymbol{\Lambda} \boldsymbol{\eta}^{\mathcal{E}} \\ &= \|\boldsymbol{\zeta}^{\mathcal{E}}\|^2. \end{aligned} \quad (14)$$

Generally, one expects $Y(\infty, \mathcal{E}) > Y(\infty)$, because the estimated $\boldsymbol{\Phi}^{\mathcal{E}}$ is flatter than its true value of $\boldsymbol{\Phi} = \mathbf{Id}_M$, with eigenvalues closer to 0. This is a classic result when estimating SNR (or STR) from an insufficient number of trials, a typical example of overfitting. As mentionned in the main text, there is no miracle cure to this problem, which should be addressed through appropriate methods of regularization and cross-validation (Hastie et al., 2009).

2.2 STR for finite neural ensembles

We now turn to the sensitivity embedded in finite subensembles \mathcal{K} from the population. The definitions of $\mathbf{A}_{\mathcal{K}}$ and $\mathbf{b}_{\mathcal{K}}$ used in eq. 11 amount to a projection from the full neural space \mathbb{R}^N to subensemble \mathcal{K} :

$$\begin{aligned}\mathbf{b}_{\mathcal{K}} &= \mathbf{P}_{\mathcal{K}} \mathbf{b}, \\ \mathbf{A}_{\mathcal{K}} &= \mathbf{P}_{\mathcal{K}} \mathbf{A} \mathbf{P}_{\mathcal{K}}^{\top},\end{aligned}$$

where $\mathbf{P}_{\mathcal{K}}$ is the $K \times N$ orthogonal projector on recorded neurons \mathcal{K} . Through the SVD decomposition in eq. 8-9, we reexpress these quantities as:

$$\begin{aligned}\mathbf{b}_{\mathcal{K}} &= \mathbf{D}_{\mathcal{K}}^{\top} \boldsymbol{\eta} \\ \mathbf{A}_{\mathcal{K}} &= \mathbf{D}_{\mathcal{K}}^{\top} \mathbf{D}_{\mathcal{K}},\end{aligned}\tag{15}$$

where

$$\mathbf{D}_{\mathcal{K}} := \boldsymbol{\Lambda} \mathbf{U}^{\top} \mathbf{P}_{\mathcal{K}}^{\top},\tag{17}$$

is our so-called *data matrix*, an $M \times K$ matrix with elements $d_i^m := \lambda_m u_i^m$. It represents the experimental data from neurons \mathcal{K} , expressed in mode space.

To compute the resulting sensitivity predicted by eq. 11, we note that through eq. 16, matrix $\mathbf{A}_{\mathcal{K}}$ has the same eigenvalues as its dual Gram matrix $\mathbf{D}_{\mathcal{K}} \mathbf{D}_{\mathcal{K}}^{\top}$, an $M \times M$ matrix with rank $d := \min(K, M)$ —generally equal to K . We introduce the (compact) SVD decomposition of this matrix:

$$\mathbf{D} \mathbf{D}^{\top} = \mathbf{X} \mathbf{T}^2 \mathbf{X}^{\top},$$

where $\mathbf{T}^2 > 0$ is a $d \times d$ diagonal matrix, and \mathbf{X} is an $M \times d$ matrix of orthogonal columns (for clarity we remove the unambiguous references to ensemble \mathcal{K}). It is shown easily that this decomposition also provides the SVD for $\mathbf{A}_{\mathcal{K}}$, in the form:

$$\mathbf{A}_{\mathcal{K}} = (\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1}) \mathbf{T}^2 (\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1})^{\top},$$

where $(\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1})$ is a $K \times d$ matrix of orthogonal columns, as required in the SVD decomposition. Thus, the (pseudo-)inverse of $\mathbf{A}_{\mathcal{K}}$ writes:

$$\mathbf{A}_{\mathcal{K}}^{-1} = (\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1}) \mathbf{T}^{-2} (\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1})^{\top}.$$

This allows to finally compute the experimental STR, from eq. 15-16:

$$\begin{aligned}Y(\mathcal{K}) &= \mathbf{b}_{\mathcal{K}}^{\top} \mathbf{A}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}} \\ &= \boldsymbol{\eta}^{\top} \mathbf{D} (\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1}) \mathbf{T}^{-2} (\mathbf{D}^{\top} \mathbf{X} \mathbf{T}^{-1})^{\top} \mathbf{D}^{\top} \boldsymbol{\eta} \\ &= \boldsymbol{\eta}^{\top} \mathbf{X} \mathbf{T}^2 \mathbf{X}^{\top} \mathbf{X} \mathbf{T}^{-4} \mathbf{X}^{\top} \mathbf{X} \mathbf{T}^2 \mathbf{X}^{\top} \boldsymbol{\eta} \\ &= \boldsymbol{\eta}^{\top} (\mathbf{X} \mathbf{X}^{\top}) \boldsymbol{\eta},\end{aligned}$$

making use of the fact that $\mathbf{X}^{\top} \mathbf{X} = \mathbf{Id}_d$. Intriguingly matrix \mathbf{T} , which describes the eigenvalues of $\mathbf{A}_{\mathcal{K}}$, disappears from the final equation. Only matrix \mathbf{X} , corresponding to the *eigenvectors* of $\mathbf{D} \mathbf{D}^{\top}$, remains in the equations. We note $\boldsymbol{\Delta}_{\mathcal{K}} := \mathbf{X}_{\mathcal{K}} \mathbf{X}_{\mathcal{K}}^{\top}$, which is nothing but the $M \times M$ orthogonal projector on $\text{Im}(\mathbf{D}_{\mathcal{K}})$. This leads to the final expression:

$$Y(\mathcal{K}) = \boldsymbol{\eta}^{\top} \boldsymbol{\Delta}_{\mathcal{K}} \boldsymbol{\eta}.\tag{18}$$

Neuron ensemble \mathcal{K} only appears through $\Delta_{\mathcal{K}}$. In particular, as soon as K is larger than the number of modes M , necessarily $\Delta_{\mathcal{K}} = \mathbf{Id}_M$, and $Y(\mathcal{K}) = Y(\infty)$: all modes are available experimentally, and sensitivity estimates saturate to their maximum value, independently of ensemble \mathcal{K} .

The whole analysis can be performed similarly assuming a finite number of measurement trials \mathcal{E} . The only difference is a modification in data matrix \mathbf{D} , to take into account the biases in mode space induced by an insufficient number of trials: $\mathbf{D}_{\mathcal{K}}^{\mathcal{E}} := (\Phi^{\mathcal{E}})^{\frac{1}{2}} \mathbf{A} \mathbf{U}^{\top} \mathbf{P}_{\mathcal{K}}^{\top}$, using the same square root of $\Phi^{\mathcal{E}}$ as in eq. 13. Similar computations lead to the final result:

$$Y(\mathcal{K}, \mathcal{E}) = (\zeta^{\mathcal{E}})^{\top} \Delta_{\mathcal{K}}^{\mathcal{E}} \zeta^{\mathcal{E}}, \quad (19)$$

which depends on experimental mode sensitivities (eq. 13) and on $\Delta_{\mathcal{K}}^{\mathcal{E}}$, the orthogonal projector on $\text{Im}(\mathbf{D}_{\mathcal{K}}^{\mathcal{E}})$, of dimension $d = \min(K, M, E)$.

2.3 Percept covariance for finite readout ensembles

Similarly to the approach above, we can express PCV signals in mode space. Since we do not model time, we only have access to the temporal average $\bar{\pi}_i := \int_{u>0} \pi_i(t_R - u) h_w(u) du$, where $\pi_i(t)$ is the full PCV curve from the main text. From eq. 9 of the main text, it falls easily that $\bar{\pi} = \mathbf{C} \mathbf{a}$. Using the optimal \mathbf{a} for readout ensemble \mathcal{K} (eq. 10, with $\mathbf{a} = \mathbf{P}_{\mathcal{K}}^{\top} \mathbf{a}_{\mathcal{K}}$ since \mathbf{a} has support on \mathcal{K}), we thus predict:

$$\bar{\pi}(\mathcal{K}) = Y(\mathcal{K})^{-1} \mathbf{C} \mathbf{P}_{\mathcal{K}}^{\top} \mathbf{A}_{\mathcal{K}}^{-1} \mathbf{b}_{\mathcal{K}},$$

which provides the value of $\bar{\pi}_i$ for every neuron i in the population (not only in ensemble \mathcal{K}). Making use of the same SVD decompositions as above, and of relationship $\mathbf{C} = \mathbf{A} - \mathbf{b} \mathbf{b}^{\top}$, we finally find:

$$\bar{\pi}(\mathcal{K}) + \mathbf{b} = Y(\mathcal{K})^{-1} \mathbf{U} \mathbf{A} \Delta_{\mathcal{K}} \boldsymbol{\eta}, \quad (20)$$

which expresses $\bar{\pi}(\mathcal{K})$ as a linear combination of mode distribution vectors \mathbf{u}^m . As \mathcal{K} tends to the full population, $\Delta_{\mathcal{K}}$ tends to \mathbf{Id}_M and we get $\bar{\pi}(\infty) = Y^{-1} \mathbf{b} - \mathbf{b} = Z^{-1} \mathbf{b}$, the prediction for choice signals in case of optimal readout (Haefner et al., 2013).

In turn, the population average for PCV is $\bar{W}(\mathcal{K}) := E_i(b_i \bar{\pi}_i(\mathcal{K}))^4$. Using eq. 20, and the general fact that $E_i(x_i y_i) = N^{-1} \mathbf{x}^{\top} \mathbf{y}$, we obtain

$$\begin{aligned} \bar{W}(\mathcal{K}) + E_i(b_i^2) &= N^{-1} \mathbf{b}^{\top} (\bar{\pi}(\mathcal{K}) + \mathbf{b}) \\ &= (N Y(\mathcal{K}))^{-1} \boldsymbol{\eta}^{\top} \mathbf{A}^2 \Delta_{\mathcal{K}} \boldsymbol{\eta}, \end{aligned} \quad (21)$$

because $\mathbf{b} = \mathbf{U} \mathbf{A} \boldsymbol{\eta}$ (eq. 8) and $\mathbf{U}^{\top} \mathbf{U} = \mathbf{Id}$. This reveals the interest of multiplying $\bar{\pi}_i$ by the corresponding tuning b_i (see discussion in main text): it allows to get rid of the unknown distribution vectors \mathbf{U} , and instead produce a quantity W which is directly related to the underlying modes' powers \mathbf{A} and sensitivities $\boldsymbol{\eta}$.

As it appears in eq. 21, we note $B^2 := E_i(b_i^2)$ the average square tuning in the population. With similar arguments as above, one shows that

$$B^2 = N^{-1} \boldsymbol{\eta}^{\top} \mathbf{A}^2 \boldsymbol{\eta} = N^{-1} \sum_{m=1}^M \eta_m^2 \lambda_m^2. \quad (22)$$

⁴which corresponds to the temporal integral $\int_{u>0} W(t_R - u | \mathcal{K}) h_w(u) du$ for the PCV curve $W(t | \mathcal{K})$ defined in the main text (eq. 16).

2.4 Behavior with K

We are now better armed to understand how sensitivity and PCV predictions vary as a function of the readout ensemble \mathcal{K} . We are mostly interested in averages of these quantities over randomly chosen ensembles \mathcal{K} of size K ; we thus use the generic notation $E_K(x) := E(x(\mathcal{K}) | \text{Card}(\mathcal{K}) = K)$. From eq. 18 we find: $E_K Y = \boldsymbol{\eta}^\top (E_K \boldsymbol{\Delta}) \boldsymbol{\eta}$.

To understand the properties of the $(M \times M)$ matrix $E_K \boldsymbol{\Delta}$, we view the $(M \times K)$ data matrix \mathbf{D}_K (eq. 17) as a collection of K random vectors \mathbf{d}_i in mode space, viewing neuron identities i as the random variable. Thus, $\boldsymbol{\Delta}_K$ is the orthogonal projector on the linear span of the K sample vectors $\{\mathbf{d}_i\}_{i \in \mathcal{K}}$. As a projector, its trace is equal to its rank, so we have $\text{Tr}(E_K \boldsymbol{\Delta}) = K$. Furthermore, since $K+1$ samples span on average more space than K samples, we are insured that $E_{K+1} \boldsymbol{\Delta} \succeq E_K \boldsymbol{\Delta}$, in the sense of positive definite matrices. Finally, intuition and numerical simulations suggest that $E_K \boldsymbol{\Delta}$ is almost diagonal. Indeed, as the various modes are linearly independent (eq. 2), there is no linear interplay between the different dimensions of \mathbf{d} across samples i : $E_i(d_i^m d_i^n) = N^{-1} \lambda_m^2 \delta^{mn}$, or equivalently

$$E_K(\mathbf{X}_K \mathbf{T}_K^2 \mathbf{X}_K^\top) = E_K(\mathbf{D}_K \mathbf{D}_K^\top) = K N^{-1} \boldsymbol{\Lambda}^2.$$

Assuming a form of independence between \mathbf{X} and \mathbf{T} , it is reasonable to suppose that $E_K(\mathbf{X}_K \mathbf{X}_K^\top) = E_K \boldsymbol{\Delta}$ is close to diagonal as well⁵.

Assuming that $E_K \boldsymbol{\Delta}$ is diagonal, we note its diagonal terms $\{\epsilon_K^m\}_{m=1 \dots M}$ and consider the resulting approximations of sensitivity (eq. 18) and mean PCV (eq. 21):

$$E_K Y \simeq \sum_{m=1}^M \epsilon_K^m \eta_m^2, \quad (23)$$

$$E_K(Y(\overline{W} + B^2)) \simeq N^{-1} \sum_{m=1}^M \epsilon_K^m \lambda_m^2 \eta_m^2. \quad (24)$$

The properties of $E_K \boldsymbol{\Delta}$ imply that $\sum_m \epsilon_K^m = K$ (trace property), and $\epsilon_{K+1}^m \geq \epsilon_K^m$ (growth property). As K augments, $\{\epsilon_K^m\}$ progressively “fills-in” the space of modes, starting from the modes with larger power λ_m . Indeed, the larger λ_m , the more often mode m appears in samples $\{\mathbf{d}_i\}$. As a useful image, we may think of the (very) rough approximation $\epsilon_K^m \simeq \mathbb{1}_{m \leq K}$: only the K first modes are revealed by a sample of K neurons. Naturally this is only a gross approximation, as can be seen easily by considering a single sample i ($K = 1$). From intuition and simulation, the true shape of $\{\epsilon_K^m\}$ (at fixed K) is a “smoothed” version of $\mathbb{1}_{m \leq K}$, and the degree of smoothing depends on the power law governing the spectrum $\{\lambda_m\}$.

With this image in mind, eq. 23 shows that the growth of sensitivity with K is linked to the progressive summation of mode sensitivities η_m^2 , starting from modes with highest power λ_m :

$$E_K Y \nearrow_K Y(\infty),$$

with a saturation as soon as all nonzero mode sensitivities η_m are revealed. Conversely, for PCV signals, we can make the rough assumption that $E_K(\overline{W} Y) \simeq E_K(\overline{W}) E_K(Y)$, in which case eq. 24 rewrites

$$E_K \overline{W} + B^2 \simeq N^{-1} \frac{\sum_{m=1}^M \epsilon_K^m \lambda_m^2 \eta_m^2}{\sum_{m=1}^M \epsilon_K^m \eta_m^2} := \left\langle \frac{\lambda_m^2}{N} \right\rangle_{m,K},$$

where each mode m contributes with a weight $\epsilon_K^m \eta_m^2$, and $E_K Y = \sum_m \epsilon_K^m \eta_m^2$ provides the normalization factor. Thus, $\langle \lambda_m^2 \rangle_{m,K}$ reflects the average power of modes with the higher sensitivity, that are already

⁵A rigorous proof might be accessible assuming a normal distribution for random vector \mathbf{d} . In the general case, small deviations from diagonality can probably occur.

revealed with K neurons. As K grows, $\{\epsilon_K^m\}$ progressively “fills-in” modes in the order of decreasing λ_m . Thus we expect $\langle \lambda_m^2 \rangle_{m,K}$ to decrease with K . Finally, as soon as $K \geq M$, we have $\{\epsilon_K^m\} = \{1\}$, and

$$\left\langle \frac{\lambda^2}{N} \right\rangle_{m,\infty} = N^{-1} \frac{\sum_{m=1}^M \lambda_m^2 \eta_m^2}{\sum_{m=1}^M \eta_m^2} = \frac{B^2}{Y(\infty)},$$

recognizing the expressions for B^2 (eq. 22) and $Y(\infty)$ (eq. 12). Since $Y^{-1} - 1 = Z^{-1}$, the predicted evolution of mean PCV signal with K follows:

$$\mathbb{E}_K \overline{W} \searrow_K \frac{B^2}{Z(\infty)} > 0.$$

\overline{W} is predicted to be positive, to decrease with increasing size K , and to saturate at its minimum value once all significant mode sensitivities η_m have been revealed—which is also the moment when sensitivity Y saturates at its maximum value (eq. 23), and corresponds to an optimal readout from the full population. The implications of these results in terms of extrapolation to large K are discussed in the main text.